Paidós
Básica

72

Eduardo Rabossi
(compilador)

Ned Block
Tyler Burge
Paul M. Churchland
Daniel Dennett
Jerry A. Fodor
William Lycan
Hilary Putnam
John Searle
Paul Smolensky
Stephen Stich
John Tienson

Filosofía de la mente y ciencia cognitiva



Eduardo Rabossi
(compilador)

Filosofía de la mente y ciencia cognitiva

Textos de:
Ned Block
Tyler Burge
Paul M. Churchland
Daniel Dennett
Jerry A. Fodor
William Lycan
Hilary Putnam
John Searle
Paul Smolensky
Stephen Stich
John Tienson



ediciones PAIDOS
Barcelona
Buenos Aires
México

El capítulo de Ned Block, «Advertisement for a Semantics for Psychology», fue publicado en *Midwest Studies in Philosophy*, X, 1986, 615-677.

«Beyond Belief», de Daniel Dennett, procede de *Thought and Object: Essays on Intentionality*, una compilación de Andrew Woodfield (1982). Traducción autorizada por Oxford University Press.

«The persistence of the attitudes», de Jerry A. Fodor, pertenece al libro *Psychosemantics*, que fue publicado en español por la Editorial Tecnos.

Cubierta de Mario Eskenazi

1.ª edición, 1995

ISBN: 84-493-0157-2
Depósito legal: B-23.385/1995

Impreso en Novagràfik, S.L.
Puigcerdà, 127 - 08018 Barcelona

Impreso en España - Printed in Spain

ÍNDICE

V. El significado y los contenidos mentales

A. Los Doppelgänger *y el anti-individualismo*

B. Mundos nocionales, semántica conceptual e individualismo

VI. Los modelos computacionales: las disputas del conexionismo y la inteligencia artificial

LISTA DE AUTORES

Ned Block. Profesor de filosofía en el Instituto Tecnológico de Massachusetts (MIT). Es compilador de *Readings in the Philosophy of Psychology* y autor de importantes trabajos sobre temas de filosofía de la mente y del lenguaje.

Tyler Burge. Profesor de Filosofía en la Universidad de California (Los Ángeles). Es autor de importantes trabajos sobre temas de filosofía de la mente y del lenguaje.

Paul M. Churchland. Profesor de Filosofía en la Universidad de California (San Diego). Es autor de *Scientific Realism and the Plasticity of Mind, Matter and Consciousness*, y *A neuro-computational perspective.*

Daniel Dennett. Profesor de Filosofía y Director del Centro de Estudios Cognitivos de la Universidad de Tufts. Es autor de *Content and Consciousness, Brainstorms, Elbow Room, The Intentional Stance* y *Consciousness Explained.*

Jerry A. Fodor. Profesor de Filosofía en la Universidad de Rutgers y en el Centro de Graduados de la Universidad de la ciudad de Nueva York. Es autor, entre otras obras, de *The Language of Thought, The Modularity of Mind, A Theory of Content, Representations* y *Psychosemantics.*

William Lycan. Profesor de Filosofía en la Universidad de Carolina del Norte. Es autor entre otras obras de *Consciousness, Judgment and Justification* y compilador de *Mind and Cognition.*

Hilary Putnam. Profesor de Filosofía en la Universidad de Harvard. Es autor, entre otras obras, de *Mind, Language and Reality, Meaning and the Moral Sciences, Representation and Reality* y *The Many Faces of Realism*.

John Searle. Profesor de Filosofía en la Universidad de California (Berkeley). Es autor de *Speech Acts, Intentionality, Minds, Brains and Science* y *The Rediscovery of the Mind*.

Paul Smolensky. Profesor de Ciencias de la Computación en la Universidad de Colorado (Boulder). Es autor de *Lectures on Connectionist Cognitive Modeling* y de importantes trabajos sobre modelización computacional de la mente.

Stephen Stich. Profesor de Filosofía en la Universidad de Rutgers. Es autor de *From Folk Psychology to Cognitive Science* y *The Fragmentation of Reason*.

John Tienson. Profesor de Filosofía en la Universidad Estatal de Memphis. Es autor de importantes trabajos sobre temas de filosofía de la mente y modelización computacional.

PREFACIO

El desarrollo de la filosofía de la mente en los últimos treinta años es espectacular: antiguos problemas han adquirido una lozanía inusitada; temas nuevos (y, por ende, nuevos problemas) han ensanchado y enriquecido el área; la masa bibliográfica es enorme, y la creatividad y la originalidad son notas frecuentes. Se trata de una etapa brillante en la historia de uno de los empeños filosóficos más antiguos y recurrentes: el de elucidar la naturaleza de los fenómenos psicológicos, indagar la índole de la mente humana y establecer los rasgos esenciales de las capacidades y de los procesos cognitivos; el empeño en desatar el "nudo del mundo", al decir de Schopenhauer.

Los problemas filosóficos —al menos los que la tradición reconoce como tales— se transmiten de una generación de filósofos a otra, a través de una dialéctica peculiar e interna. Los problemas centrales de la filosofía de la mente no son una excepción a ese tipo de transmisión y elaboración secular. Pero sería erróneo atribuir el *boom* actual únicamente a la dinámica interna del pensamiento filosófico. En realidad, el impulso más fuerte es extrafilosófico. Proviene del ámbito científico: de la neurociencia, la biología y la ciencia cognitiva. Filosofar acerca de lo mental implica, hoy día, involucrarse en una práctica teórica relacionada de maneras diversas con la teorización de biólogos, lingüistas, psicólogos, expertos en Inteligencia Artificial, psiquiatras biológicos, expertos en sistemas computacionales. Con alguna frecuencia, lo inverso también se da. En los niveles más elevados de cada una de esas disciplinas suelen surgir problemas y/o formularse hipótesis de indudable raigambre filosófica. Ese es un incentivo adicional para el filósofo de la mente.

Esta compilación supone, pues, un universo heterogéneo de lectores. Está dirigida, naturalmente, a filósofos: a los no iniciados en estos

temas, y a los ya iniciados que se propongan utilizarla en cursos y seminarios. También está dirigida a los científicos: a los que han tropezado en su práctica teórica con cuestiones del talante de las que aquí se ventilan, y a los que simplemente tengan curiosidad por conocer la trastienda filosófica de su propio quehacer. Está dirigida, por fin, a lectores que sientan curiosidad por conocer qué pasa en esta área de la filosofía.

Decidir qué trabajos van a integrar una compilación siempre es un acto discrecional. Esa discrecionalidad aumenta cuando la compilación es de temática abierta y la oferta de trabajos disponibles es enorme. Los trabajos seleccionados se ocupan, principalmente, de problemas relacionados con el diseño de modelos filosóficos viables de los fenómenos mentales, en su relación con la psicología de sentido común y las teorías científicas (secciones II y III), la índole de las actitudes proposicionales y la tesis del lenguaje del pensamiento (sección IV), el papel del significado en relación con los contenidos mentales (la representación mental) (sección V) y la estructura y viabilidad de los modelos computacionales de la mente (sección VI). Muchos temas importantes están excluidos. Por ejemplo, cuestiones ontológicas como la causación mental, el reduccionismo, la superveniencia, y "cuestiones crónicas" como el *status* y la naturaleza de los estados de conciencia. Es de desear que puedan ser tema de futuras compilaciones. Con el fin de ubicar al lector en los problemas que trata la compilación y de ofrecerle una bibliografía adicional, incluyo en la Sección I un trabajo introductorio. He compuesto además un Índice Temático que, espero, ayudará a "cruzar" material de un trabajo a otro(s) o a "seguir" un tema a través de distintos trabajos.

La idea de la compilación tomó forma durante mi residencia en la Universidad de California, Berkeley, como Fulbright Scholar; la desarrollé como parte de mis tareas en el Consejo Nacional de Investigaciones Científicas y Técnicas y en el seno del Proyecto de Investigación "Representaciones, significado e intencionalidad" (UBACYT FIL-064), que dirijo, y culminó en el National Humanities Center. Agradezco a la Fulbright Commission, al CONICET, a la Secretaría de Ciencia y Técnica de la Universidad de Buenos Aires y al National Humanities Center el apoyo brindado a mi actividad filosófica.

Debo a Donald Davidson, Daniel Dennett y Ernesto Sosa importantes sugerencias iniciales. Todos los autores incluidos respondieron a mis requerimientos con generosidad. Ned Block, Paul Churchland, Jerry Fodor, William Lycan y John Tienson ayudaron a traducir algún

párrafo rebelde, facilitaron trámites formales y/o formularon sugerencias acerca del contenido de la compilación. En el National Humanities Center, Ruth Marcus y Alfred Mele fueron consultores pacientes. Allí, Karen Carroll y Lynda Morgan pasaron en limpio manuscritos de dudosa lectura. A todos, muchas gracias.

Las traducciones han estado a cargo de miembros del Proyecto de Investigación mencionado. Al encararlas, han intentado producir un material técnicamente confiable, susceptible de ser empleado en cursos y seminarios de exigencia profesional. Se ha cuidado la corrección conceptual, y para dar al lector un control del texto, se han agregado entre corchetes las expresiones técnicas inglesas y ciertas expresiones coloquiales de traducción ambigua. Los trabajos se publican respetando la diagramación elegida por cada autor, y su manera de ubicar y citar la bibliografía. Agradezco de manera muy especial la responsabilidad y el entusiasmo de los miembros del Proyecto. Lograron que la tarea de traducir se transformara a menudo en un ejercicio filosófico fructífero. Sus nombres se leen al pie de las traducciones, como corresponde.

Agradezco a la profesora Cristina González la generosa ayuda que prestó en el tramo final de la edición. En las estimables personas de la señora Marita Gottheil y el profesor Manuel Cruz, agradezco a la Editorial Paidós haber hecho posible esta publicación.

I

LA FILOSOFÍA DE LA MENTE Y LA CIENCIA COGNITIVA

CAPÍTULO 1

CÓMO EXPLICAR LO MENTAL: CUESTIONES FILOSÓFICAS Y MARCOS CIENTÍFICOS

Eduardo Rabossi

1. Acerca de la filosofía de la mente

Según cierta concepción de la filosofía de la mente —la concepción tradicional—, filosofar acerca de los fenómenos mentales es llevar a cabo una reflexión a priori sobre los conceptos mentales, articular qué está involucrado en el contenido de tales conceptos y descubrir verdades necesarias acerca de ellos y de la mente. No se requiere, en principio, adquirir conocimiento empírico de los fenómenos y las operaciones mentales. El filósofo de la mente accede al ámbito conceptual que le es propio y trabaja en él. Esto implica suponer que la naturaleza esencial de los fenómenos mentales está contenida, de alguna manera, en los conceptos mentales. El objetivo de esa práctica será articular un conjunto de verdades que valgan para cualquier ejemplificación concebible de los fenómenos mentales. El filósofo de la mente (como todo filósofo que se precie de tal) es un especialista en descubrir los límites de lo conceptualmente posible.

Ésta es una descripción de la concepción tradicional de la filosofía de la mente "en estado puro". En la práctica, este formato ha sido llenado de maneras muy diversas: enfatizando más o menos la pertinencia del plano lingüístico, prestando mayor o menor atención a las convicciones de sentido común o a cierta información empírica, centrando el análisis en un concepto o en una familia de conceptos, enfatizando o no los rasgos públicos de los fenómenos mentales, etcétera. Lo producido como filosofía de la mente en el mundo anglosajón durante las décadas del 50 y el 60, es una versión del formato básico [Ryle, 1949; Chappell, 1962; Guftafson, 1964; Hampshire, 1966]. Pero la concepción tradicional no es una cosa del pasado. Ciertas versiones puras cuentan con defensores explícitos [McGinn, 1982a]. Y versiones menos puras cuentan con

muchos partidarios implícitos. Téngase presente que el análisis filosófico es un tipo de análisis conceptual de carácter a priori.

Hay filósofos que se oponen a esa manera de concebir la filosofía de la mente y le critican, con razón, el monopolio que atribuye a la reflexión a priori y el aislamiento que impone respecto del saber científico. La filosofía de la mente, sostienen, debe estar estrechamente relacionada con la ciencia. Para algunos, los problemas de la filosofía de la mente tienen un carácter empírico: las tesis filosóficas poseen un estado conjetural que únicamente puede ser resuelto por ciertos programas de investigación científica. Para otros, la filosofía tiene que cumplir el rol de un agente especulativo y sintetizador respecto de ciertos programas científicos pertinentes: los de la psicobiología [Bunge, 1977, 1980], la neurociencia [Churchland, capítulo 2 de esta compilación; Churchland, 1986] y la inteligencia artificial [Boden, 1977, 1990; Pylyshyn, 1984]. La psicología cognitiva posee un atractivo especial. Hay filósofos de la mente que describen su quehacer, precisamente, como filosofía de la psicología. Y cuando usan el rótulo con estrictez, hacen referencia a una reflexión de segundo nivel que tiene como objeto describir, articular y criticar las presuposiciones teóricas de la psicología, sin excluir algunos problemas tradicionales de la filosofía de la mente [Block, 1980b; Block y Alston, 1984; Bunge y Ardila, 1987]. Pese a estas y otras diferencias (que no son entre sí excluyentes), el objetivo común es claro y plausible: filosofar concurrentemente con ciertos programas científicos de punta, poner un límite al carácter a priori atribuido a la reflexión filosófica y no excluir ciertos problemas tradicionales de la filosofía de la mente.

Dada la índole de esta compilación, "filosofía de la mente" se refiere a *todas* las modalidades referidas.

Las discrepancias acerca de la índole de la filosofía de la mente son promovidas por una situación comparable a la que se produjo en los siglos XVII y XVIII con el advenimiento de la ciencia físico-matemática. En las últimas décadas, la investigación de los fenómenos y procesos cerebrales y cognitivos ha experimentado un cambio revolucionario, que ofrece maneras inéditas de conceptualizarlos y estudiarlos. Es natural que ese fenómeno produjera un fuerte impacto en la filosofía de la mente. Echemos, pues, una rápida mirada a las nuevas ciencias de la mente.

2. Las nuevas ciencias de la mente y el giro cognitivo

Howard Gardner, en su clásico estudio sobre la "revolución cognitiva" [Gardner, 1985], distingue los *inputs* teóricos que le sirvieron de base, los "encuentros catalíticos" que la engendraron, y su desarrollo posterior. Cada una de esas etapas y el proceso global que componen, comparten un rasgo curioso. Científicos que provienen de disciplinas diferentes, efectúan descubrimientos que tienen un efecto común: generan una manera novedosa de modelizar y de investigar los fenómenos cerebrales y cognitivos.

Los *inputs* teóricos son variados. La lógica matemática; los aportes de Alan Turing (la descripción formal de una máquina que pueda llevar a cabo todo cálculo concebible y la elaboración de un test que hace imposible discriminar las respuestas de una máquina adecuadamente programada, de las de una persona); las contribuciones de John von Neumann (en especial, la descripción formal de una máquina de Turing que pueda preparar y elaborar sus propios programas); la conjetura de Alonzo Church (todo procedimiento susceptible de ser derivado de manera explícita puede ser computado con funciones recursivas); la cibernética o teoría de los sistemas autorregulados (Wiener); la teoría matemática de la información (Shannon); la modelización lógica de las neuronas y las redes neurales (McCullock y Pitts); la teoría acerca del ensamblaje de las neuronas y su relación con los patrones de comportamiento (Hebb); los estudios realizados en laboratorios de Berlín, Moscú, Nueva York, Oxford y París sobre síndromes neuropsicológicos de víctimas de la guerra. Este enorme caudal teórico influyó en los trabajos que comenzaron a producirse hacia fines de la década del 40, coincidiendo con la aparición de los primeros computadores electrónicos.

Una serie de *simposia* alimentados por los aportes de una generación talentosa dio contenido definitivo al proceso. Chomsky, Newell, Simon, McCarthy, Miller, Bruner, Pribram, Gallanter, Minsky, son algunos representantes notables.

A partir de los años 70 la expresión 'ciencia cognitiva' comienza a ser usada para designar, no sin ambigüedad, un nuevo ámbito de investigación de la mente humana, con aportes de la lingüística, la psicología, la inteligencia artificial, la neurociencia y la filosofía.

Pero, ¿qué es la ciencia cognitiva? No es, por cierto, una disciplina científica como la física, la biología o la antropología. ¿Qué es, entonces? Una pintoresca caracterización debida a Flanagan es, quizá, la respuesta más realista. La ciencia cognitiva es un comité bien organizado

y creciente de disciplinas y sub-disciplinas, todas las cuales alegan que pueden contribuir a nuestra comprensión de lo mental. Específicamente, la ciencia cognitiva es una confederación [formada por] la filosofía (en particular, la filosofía de la mente, la filosofía del lenguaje, la gnoseología y la lógica), la psicología cognitiva, la neurociencia, la lingüística y la ciencia de la computación [Flanagan, 1984].

¿Sobre qué bases se asienta esa atípica confederación de disciplinas científicas? Lo que sigue intenta reconstruir su matriz teórica mínima.

1. Los seres humanos y, en general, todo ingenio al que se le atribuyen estados y procesos cognitivos, son sistemas procesadores de información. 'Información' hace referencia a unos ítemes abstractos sobre los que se opera, y 'procesamiento' hace referencia a secuencias o series ordenadas de operaciones.

2. El procesamiento de información involucra reglas, elementos simbólicos con propiedades sintácticas (formales) y operaciones computacionales (algorítmicas) sobre esos ítemes.

3. Todo proceso cognitivo involucra procesamiento de información.

4. Los elementos simbólicos tienen un carácter representacional; las representaciones internas son de índole "descripcional" (proposicional), aunque no se excluyen, en principio, representaciones de índole "pictórica" (imágenes).

5. El estudio de los mecanismos cognitivos exige un nivel abstracto de análisis, es decir, un nivel que permita especificar el método a través del cual el organismo o ingenio lleva a cabo su función informacional.

6. Ese nivel abstracto es el nivel computacional (*software*); todo proceso cognitivo es un proceso computacional.

7. Todo proceso cognitivo se implementa en una base física (*hardware*), pero la especificación computacional subdetermina el nivel físico de implementación, en el sentido de que bases físicas diferentes pueden implementar un mismo programa.

Esta base teórica mínima y su aplicación disciplinaria tiene una

obvia relevancia filosófica y da origen, además, a problemas filosóficos. De ahí el *giro cognitivo* que ha dado la filosofía de la mente en las últimas dos décadas. Ese giro involucra una serie de rasgos peculiares. Sólo indicaré dos. Piénsese en la significación que tiene la computación electrónica. Hasta hace unas pocas décadas la tecnología (casi sin excepción) sólo había provisto de ingenios que nos permitían ampliar las posibilidades físicas. La computación electrónica es un avance tecnológico radicalmente distinto. *No sólo amplía enormemente nuestras aptitudes cognitivas, sino que permite emularlas y modelizarlas en sus propios términos.* Ambos rasgos son absolutamente novedosos y de consecuencias inimaginables. En segundo lugar, la ciencia cognitiva no es filosóficamente neutral. Implica el rechazo del "paradigma" conductista y la consiguiente adopción de "paradigmas" teóricos que, al margen de diferencias específicas, involucran la adopción de un enfoque mentalista (los estados/procesos psicológicos son estados *internos* de un sistema [Nudler, 1975]). No debe sorprender que Descartes, para algunos, y Kant, para otros, sean antecesores reverenciables y hasta emulables.

La filosofía de la mente debe "leerse" hoy sobre este trasfondo teórico. Por eso es una empresa fascinante.

[Sobre el desarrollo histórico y la significación de la ciencia cognitiva, véase Gardner, 1985 y Meyering, 1989. Sobre los fundamentos y el contenido de la ciencia cognitiva, véase Stillings, 1987; Haugeland, 1981. Sobre la filosofía de la mente y de la cognición, véase Block, 1980a; Bechtel, 1988; Churchland, 1988; Lycan, 1990; Beakley y Ludlow, 1992. Sobre Inteligencia Artificial, véase Simon, 1969; Haugeland, 1985. Sobre lingüística, véase Chomsky, 1966, 1979. Sobre psicología cognitiva, véase Lindsay y Norman, 1983; de Vega, 1987. Sobre neurociencia, véase Globus, Maxwell y Sovodnik, 1976; Churchland, 1986.]

Hasta aquí los comentarios generales acerca de la filosofía de la mente, el carácter de la revolución de las ciencias de la mente y el giro cognitivo. En lo que sigue describiré el marco teórico en el que se inserta cada una de las secciones de esta compilación.

3. El status *de la psicología de sentido común*

Los humanos somos entidades concientes de sí mismas, capaces de experimentar distintos tipos de estados psicológicos y de producir comportamiento inteligente. Percibimos, categorizamos y representamos de maneras diversas el mundo físico, interpersonal y cultural en el que esta-

mos inmersos. Actuamos persiguiendo propósitos, por lo general, definidos. Expresamos nuestros estados psicológicos mediantes gestos, ademanes y, sobre todo, mediante expresiones lingüísticas. Todo lenguaje natural tiene una parte "mentalista" que permite expresar y describir los estados mentales propios y adscribir estados mentales a los demás.

Creemos que las acciones reflejan sucesos mentales concomitantes (qué se cree, desea, piensa, siente; cuáles son las motivaciones, los propósitos). Comprendemos, interpretamos, explicamos y predecimos las acciones de los demás mediante la adscripción de estados mentales (básicamente, de deseos y creencias). Esa capacidad de predecir y explicar la conducta suele tener éxito. Y es importante que lo tenga: es una condición necesaria para que las relaciones interpersonales y el mundo social sean posibles. No aprendemos todo esto de manera explícita. La adquisición y el desarrollo de estas aptitudes forma parte del proceso de aculturación.

Casi sin excepción, los filósofos y los psicólogos han presupuesto ese marco cotidiano como base para sus reflexiones e investigaciones. El análisis de su valor cognoscitivo y ontológico sólo se dio de manera tangencial en las discusiones sobre el sentido común [Rabossi, 1979]. Pero en los últimos diez años, el carácter y *status* de la psicología de sentido común (*folk psychology*, psicología intencional, psicología de las actitudes proposicionales, psicología popular) ha ocupado un lugar importante en la agenda de la filosofía de la mente.

Se suele sostener que la psicología de sentido común es una teoría [Sellars, 1958; Feyerabend, 1963]. El argumento es simple. El vocabulario de la psicología de sentido común hace referencia a estados y procesos mentales y supone, en consecuencia, que tales estados y procesos existen. La atribución de esos estados y procesos permite interpretar, explicar y predecir las acciones de los demás. En tal caso, algunos de los términos empleados desempeñan el papel de términos teóricos, y las explicaciones y predicciones suponen la existencia de regularidades y leyes. Todos éstos son los rasgos característicos de una teoría.

El argumento, aceptado por la mayoría de los filósofos de la mente, tiene consecuencias importantes. Admitir que la psicología de sentido común es una teoría, autoriza a evaluarla con los patrones que se aplican a cualquier teoría y a preguntar por su integración con otras teorías. Es posible, entonces, que la psicología de sentido común resulte falsa, que no existan los estados y procesos de los que habla. Es posible eliminarla y, en consecuencia, reemplazarla por una teoría mejor.

En "El materialismo eliminativo y las actitudes proposicionales"

(capítulo 2 de esta compilación), Paul M. Churchland adopta una posición radical. Sostiene que la psicología de sentido común es falsa y concluye que debe ser eliminada y sustituida por una teoría basada en la neurociencia. Argumenta que la psicología de sentido común es una teoría anquilosada que no permite explicar y solucionar importantes cuestiones psicológicas. Está incapacitada, además, para integrar una visión científica global de los seres humanos. Y como (probablemente) sus términos carecen de referencia, una ontología adecuada no puede incluir estados mentales. En compensación, Churchland ofrece varios ejemplos hipotéticos del tipo de conocimiento que podría proporcionar una neurociencia completa. [Véase Churchland, 1979, 1988; para sus argumentos recientes y su giro al conexionismo véase Churchland, 1991.]

Hay versiones menos drásticas que describen ciertas peculiaridades del funcionamiento de la psicología de sentido común y concluyen que la tornan incompatible con una ciencia madura de la mente. Las características de las entidades involucradas en las atribuciones de la psicología de sentido común y los mecanismos explicativos e individuativos a los que apela, no encajan ni en la ontología ni en los marcos causal-explicativos de una ciencia cognitiva adecuadamente desarrollada. [Véase Stich, 1978, 1982, 1983; para sus argumentos recientes y el giro al conexionismo, véase Ramsay, Stich y Garon, 1991.]

Estos argumentos eliminativistas han sido respondidos de distintas maneras [véase, entre otros, Horgan y Woodward, 1985; Baker, 1985; Jackson y Petit, 1990]. Todos los "defensores" de la psicología de sentido común son *realistas*, en algún sentido pertinente, respecto de los estados y procesos mentales.

Jerry Fodor es el más persistente defensor del realismo intencional. Las actitudes proposicionales (los estados mentales con un contenido proposicional) cumplen un rol esencial en las explicaciones psicológicas de sentido común, poseen una existencia real, son semánticamente evaluables y causalmente eficaces. La ciencia cognitiva reconoce y desarrolla esos rasgos y mecanismos. El realismo de Fodor tiene una importante razón de ser. La decisión acerca del carácter real o no real de las actitudes proposicionales, define la concepción teórica de las representaciones mentales. Ello equivale a definir, en gran medida, el modelo de la mente [Fodor, 1985, 1987]. En "La persistencia de las actitudes" (capítulo 3 de esta compilación), Fodor expone los sutiles y eficaces mecanismos de las explicaciones psicológicas de sentido común, analiza el papel y la forma de las regularidades presupuestas, estudia las propiedades esenciales de las actitudes

proposicionales y delinea los rasgos básicos de la teoría representacional de la mente (ver la sección V).

Todas estas posiciones son contrapuestas por un enfoque instrumentalista de las actitudes proposicionales. Daniel Dennett, su inspirador, admite que muchas de las convicciones implícitas en la psicología de sentido común son vulnerables a los avances científicos, pero reconoce que poseen un gran poder predictivo. Coincide con los realistas en que las personas tienen deseos y creencias aunque discrepa con ellos en dos puntos importantes: 1) las creencias y los deseos no son estados discretos de un sistema conductual-causal, y 2) su *status* es equiparable al de los centros de gravedad o al del Ecuador (son, en la terminología de Reichenbach, *abstracta*). [Véase Dennett, 1978, 1987, 1991.]

Pero cabe aún otra opción: negar que la psicología de sentido común sea una teoría, en un sentido adecuado del término. El desafío consiste entonces en elaborar un enfoque que tenga aptitud descriptivo-explicativa, que garantice la independencia de los problemas pertinentes de las ciencias de la mente y que sea compatible con el contenido y el desarrollo de dichas ciencias. [Véase von Eckardt, 1984; Kitcher, 1984; Baker, 1987; McDonough, 1991; Pérez, 1993.] [Sobre psicología de sentido común véase, además, CNRS, 1988; Greenwood, 1991, y el número especial de *Mind and Language*, 1993, 8.]

4. El funcionalismo y la naturaleza de los fenómenos mentales

La polémica acerca del *status* de la psicología de sentido común se enlaza con un problema filosófico recurrente: el de la naturaleza de los fenómenos mentales. Existe la convicción de que un enfoque funcionalista permite superar las dificultades que han afectado al dualismo substancialista, al conductismo y a la teoría de la identidad mente/cerebro [sobre estas teorías véase Campbell, 1970; Churchland, 1988; Bechtel, 1988].

Los funcionalistas sostienen las siguientes tesis básicas: 1) la naturaleza de un estado mental es su rol causal, explicitable en términos de *inputs* sensoriales, otros estados mentales internos y *outputs* fisiológico-conductuales; 2) la función (el rol causal) es distinta de la implementación física (el ocupante del rol); 3) la explicitación del rol causal permite identificar tipos (clases, propiedades) funcionales; la implementación física involucra en cambio, casos (ejemplares) específicos de esos tipos en el nivel estructural; 4) los casos de los tipos funcionales son idénticos

a estados (caso) físicos; y 5) los tipos funcionales son independientes de la base de implementación, es decir, son realizables en un número indefinido de bases de implementación (argumento de la realizabilidad variable). Estas tesis permiten sostener que lo mental no es reducible a lo físico y que la psicología es una ciencia autónoma. Pero, al mismo tiempo, permiten adherir al fisicalismo, esto es, la tesis de que la base de implementación de los fenómenos mentales es de naturaleza física. Estos son, sin duda, rasgos atractivos del funcionalismo.

Si se tienen presentes los puntos 5, 6 y 7 de la base teórica de la ciencia cognitiva, se advertirá la íntima relación que el funcionalismo guarda con ellas. Las primeras propuestas funcionalistas coinciden con el desarrollo de la ciencia cognitiva, y la concepción de la mente como un tipo de sistema funcional está moldeada por la "metáfora del computador", la distinción *software/hardware* y la independencia del primero respecto del segundo.

En "Las dificultades del funcionalismo" (capítulo 4 de esta compilación), Ned Block ofrece una excelente descripción del funcionalismo y de sus variantes (1.0 y 1.1). En el resto del trabajo señala varias dificultades que afectan a toda concepción funcionalista. La tesis de que la mente es un sistema funcional no es, en realidad, una tesis substantiva. Es posible imaginar sistemas que satisfacen los requerimientos funcionalistas pero que, obviamente, carecen de estados mentales (los robots de cabeza homuncular, el sistema económico de un país, por ejemplo). El argumento se extiende a los *qualia* (los contenidos cualitativos de ciertos estados mentales) y a las caracterizaciones no restringidas de los *inputs* y *outputs*. El funcionalismo resulta ser demasiado "liberal". Paralelamente, Block señala el carácter "chauvinista" (restringido a lo humano) de algunas caracterizaciones funcionalistas. Las versiones canónicas del funcionalismo no pueden lograr un punto de equilibrio entre el liberalismo y el chauvinismo.

En "El materialismo eliminativo y las actitudes proposicionales" (capítulo 2, § 4, de esta compilación), Churchland formula una crítica al funcionalismo que coincide, en parte, con la objeción "liberal" de Block, pero que va más lejos: la funcionalización de una teoría puede traer aparejada su irrefutabilidad. El ejemplo de la versión funcional de la alquimia es, sin duda, revelador.

En "La continuidad de niveles en la naturaleza" (capítulo 5 de esta compilación) William Lycan ofrece una versión del llamado 'funcionalismo teleológico', que permitiría superar esas y otras limitaciones del funcionalismo canónico. Lycan rechaza la aplicación rígida, en dos nive-

les, de la distinción *software/hardware*: el nivel computacional abstracto (que corresponde al programa, a la descripción psicológica) y el nivel estructural concreto (que corresponde a la descripción biológica, fisiológica, del organismo). Postula, en cambio, una organización jerárquicamente ordenada de múltiples niveles, en la que la distinción es relativa a cada nivel organizacional [Wimsatt, 1976]. Adopta como posición metodológica el funcionalismo homuncular [Dennett, 1978, 1987], que interpreta como la tesis de que las personas somos entidades corporativas que "llevamos a cabo, corporativamente, muchas funciones inmensamente complejas". La mente puede ser descompuesta en un número de homúnculos que, a su vez, pueden ser descompuestos en sub-homúnculos hasta arribar a un nivel en el que las tareas son tan simples que cualquier ingenio mecánico puede realizarlas ("el ejército de idiotas" de que habla Dennett). El método es tomado de la Inteligencia Artificial [Simon, 1969]. Lycan impone, entonces, un requisito teleológico: la implementación física de una adscripción teleológica vale si el organismo tiene una "integridad orgánica genuina", y el estado atribuido tiene *para* el organismo un rol funcional. Sobre esta base, Lycan responde a las críticas de Block. [Sobre el funcionalismo teleológico, véase Millikan, 1984; Sober, 1985; Millikan, 1986; Dretske, 1988; Millikan, 1989.] [Sobre el funcionalismo en general, véase Bechtel, 1988; Block, 1980a; Lycan, 1990, y la bibliografía que citan.]

5. *Las actitudes proposicionales, la teoría representacional de la mente y la hipótesis del lenguaje del pensamiento*

Aceptar el funcionalismo implica reconocer que existe un nivel explicativo válido para la ciencia cognitiva, insertado entre el nivel de las explicaciones de la psicología de sentido común y el de las explicaciones neurales. Pero aceptar el funcionalismo no implica que se esté en condiciones de explicar cómo la mente representa a la realidad y en virtud de qué los estados psicológicos poseen un contenido y desempeñan un rol causal. La teoría representacional de la mente (TRM) y la hipótesis del lenguaje del pensamiento (HLP) se proponen sentar las bases de esa explicación.

La TRM sostiene, básicamente, que los estados mentales (de manera especial, las actitudes proposicionales) son estados representacionales internos y que la actividad mental consiste en adquirir, transformar y usar información [Sterelny, 1990]. La tesis acerca del carácter represen-

tacional de los estados mentales está directamente conectada con la doctrina clásica de la intencionalidad: ciertos estados mentales tienen un contenido al que "están dirigidos" o son "acerca de" lo que ese contenido expresa.

¿Qué involucra, concretamente, el carácter representacional? ¿Cuál es el vehículo propio de la actividad representacional? La HLP sostiene que los contenidos de las creencias y de otras actitudes proposicionales poseen un carácter lingüístico. Son secuencias compuestas, con estructura sintáctica, que hacen referencia a propiedades en el mundo. Su significado está determinado por las propiedades semánticas de las partes y por las reglas gramaticales asociadas a las estructuras sintácticas. Tienen condiciones de verdad y guardan entre sí relaciones lógicas de implicación. Los computadores emplean un "lenguaje de máquina" y la mente humana emplea, en un sentido similar, un "lenguaje del pensamiento". Por otra parte, hay estados físicos que hacen las veces de los elementos de un vocabulario y que implementan las reglas que combinan esos elementos. Las configuraciones producidas poseen los contenidos que la psicología de sentido común atribuye a las actitudes proposicionales [Lycan, 1988, 1990]. La HLP ha sido presentada en versiones diferentes [Sellars, 1963; Harman, 1975; Fodor, 1975; Field, 1978; entre otros. También ha sido criticada desde distintos puntos de vista, véase Dennett, 1978; Churchland, 1986; Schiffer, 1987, entre otros].

En *The Language of Thought* (1975) y en una serie de trabajos posteriores, Fodor ha desarrollado una influyente y discutida doctrina que combina la TRM y la HLP con un enfoque computacional de la mente (TCM). En "La persistencia de las actitudes" (capítulo 3 de esta compilación), sostiene que la TRM es la única propuesta viable acerca del funcionamiento de la psicología del sentido común y que es, además, la base teórica de la psicología científica. Fodor argumenta que tener una actitud proposicional (creer que *p*, por ejemplo) es tener ejemplificado en la cabeza un símbolo mental del "mentalés" que significa que *p*. La TCM completa el cuadro: las diferentes actitudes proposicionales (creer, suponer, desear) son relaciones computacionales. La metáfora del computador desempeña, además, un papel heurístico: la computación nos muestra cómo se conectan a través de la sintaxis, las propiedades causales de un símbolo con sus propiedades semánticas, y cómo la sintaxis de un símbolo puede determinar las causas y los efectos de sus casos. Finalmente, Fodor defiende a la TRM de dos objeciones: puede haber actitudes proposicionales sin representaciones mentales y representaciones mentales sin actitudes proposicionales [véase Dennett, 1978, 1987].

En "Las actitudes proposicionales" (capítulo 6 de esta compilación) Fodor describe un conjunto de condiciones a priori que debe satisfacer una teoría adecuada de las actitudes proposicionales e identifica lo que da en llamar "las tres piezas en juego": los adscriptores de creencias (la oración 'Juan cree que está lloviendo'), los complementos de los adscriptores de creencias (la frase 'Está lloviendo') y las (oraciones) correspondientes a los adscriptores de creencias (la oración del lenguaje público 'Está lloviendo'). Los primeros son verdaderos en virtud de relaciones de carácter funcional/causal entre los organismos y los casos de las fórmulas que correspondan. Los complementos efectúan la conexión con la fórmula interna y con la (oración) correspondiente. Fodor concluye advirtiendo que el problema básico que tiene que resolver la TRM es cómo se conectan esas relaciones internas con el mundo, qué significa sostener que un sistema de relaciones internas está interpretado semánticamente.

En "La teoría sintáctica de la mente" (capítulo 7 de esta compilación) Stephen Stich critica a la TRM. Stich llama TRM Fuerte a la teoría de Fodor, y le atribuye dos tesis básicas: que una ciencia cognitiva seria se funda en la expectativa de que las generalizaciones de la psicología de sentido común puedan ser sistematizadas, y que esas generalizaciones hacen referencia a los contenidos de los estados mentales. Stich sostiene que formular las generalizaciones cognitivas en términos de contenido implica excluir generalizaciones significativas y caer en una vaguedad endémica, producto de las dificultades que genera la dependencia contextual y el esquema individuativo adoptado. La teoría sintáctica de la mente (TSM) sostiene, en cambio, que los estados cognitivos son mapeados en objetos sintácticos abstractos, de modo que las interacciones de los estados cognitivos y los nexos causales con los estímulos y la conducta, son descriptos en términos de las propiedades y relaciones sintácticas de tales objetos abstractos. En definitiva, la TSM es un mejor candidato para el científico cognitivo que la TRM Fuerte, porque elimina los contenidos como intermediarios. La argumentación de Stich incluye la evaluación de dos principios acerca de los presupuestos de las teorías psicológicas: el solipsismo metodológico y el principio de autonomía.

Putnam señaló, hace años, que los filósofos tradicionales han sido solipsistas metodológicos, es decir, que han dado por supuesto que los estados psicológicos propiamente dichos sólo involucran la existencia del individuo al que se los adscribe. Ello implica adoptar un programa "estrecho" [*narrow*] para la psicología y postular estados psicológicos

estrechos (como distinto de postular estados psicológicos "amplios" [*wide*]) [Putnam, 1975 b]. Fodor ha defendido expresamente el solipsismo metodológico, en tanto estrategia investigativa que restringe la psicología cognitiva a la postulación de estados mentales sobre los que se realizan operaciones formales [Fodor, 1980]. Stich argumenta que la TRM Fuerte no condice con el solipsismo metodológico porque, de un lado, caracteriza a los estados mentales en términos de contenido, de propiedades semánticas, y del otro lado, no parece tener un buen argumento para sostener que las propiedades semánticas no desempeñan ningún papel en la especificación de las generalizaciones psicológicas. En ese sentido la TSM congenia mejor con el solipsismo metodológico. Sin embargo, Stich considera preferible el principio de autonomía: el psicólogo sólo debe considerar estados y procesos que supervengan en los estados físicos internos y corrientes del organismo; es decir, debe ignorar toda diferencia entre organismos que no surja de diferencias de los estados físicos pertinentes. Esto hace que la "historia" y el entorno de los organismos sean irrelevantes para la teoría psicológica, en tanto no influyan en tales estados físicos [véase Stich, 1978].

6. El significado y los contenidos mentales

Una posición de corte fodoriano acerca del contenido, puede sintetizarse así: 1) en la psicología de sentido común y en la psicología científica se debe apelar a una noción estrecha del contenido (las actitudes proposicionales se individuan en función de sus propiedades intrínsecas, es decir, no relacionales); 2) los estados cerebrales se individuan de la misma manera; 3) los estados mentales supervienen en los estados cerebrales; 4) las propiedades semánticas y los roles causales dependen de las configuraciones sintácticas; y 5) las propiedades relacionales (las propiedades que corresponden al contenido amplio de los estados mentales) son pertinentes para su carácter representacional, es decir, para las conexiones con el entorno no mental y comunitario; no son pertinentes, en cambio, para la individuación y para el contenido de los estados mentales. Ésta es, típicamente, una posición individualista respecto de los contenidos mentales. ¿Es el individualismo una posición viable?

En "Significado y referencia" (capítulo 8 de esta compilación) Hilary Putnam presenta un famoso "experimento mental" [véase también Putnam, 1975b]. Propone imaginar un planeta (la Tierra Gemela) en el que en vez de H_2O hay XYZ. Ambas substancias son fenoméni-

camente indistinguibles. En ese planeta hay un gemelo que es idéntico a uno, molécula por molécula. Si yo y mi gemelo decimos 'El agua calma la sed', la oración no significa lo mismo. En mi caso se refiere a H_2O, en el de mi gemelo, a XYZ. Esto muestra que la referencia y, en general, el significado de los términos de clases naturales, no depende ni de su relación con otros términos ni de los estados internos de los hablantes (esto último es lo que debería sostener el individualista). Putnam concluye que los significados "no están en la cabeza". Su argumento apunta, en realidad, contra el intento de explicar el significado de las palabras de clases naturales en términos de descripciones y forma parte de la elaboración de teorías causales de la referencia [Kripke, 1980]. Pero el argumento implica, al menos, que el contenido de los estados mentales depende de sus propiedades relacionales, es decir: 1) que los orígenes causales de los estados psicológicos desempeñan un papel crucial en la determinación del contenido representacional, y 2) que los estados mentales no supervienen en los estados neurales. El argumento implica la reivindicación de una noción amplia de contenido.

En "El individualismo y la psicología" (capítulo 9 de esta compilación), Tyler Burge se opone, a su vez, a que se fijen restricciones a priori sobre la psicología científica, tal como resulta de la estrategia individualista, y sostiene que de hecho la psicología no opera sobre bases individualistas. Burge analiza el siguiente argumento en favor del individualismo: 1) en los experimentos mentales la conducta de los participantes es idéntica; 2) la psicología es la ciencia de la conducta; 3) dados 1 y 2, la ciencia de la conducta deberá dar en ambos casos las mismas explicaciones y predicciones; y 4) no hay espacio, pues, para explicar la conducta en términos de aquello a lo que los respectivos estados mentales hacen referencia. Burge rechaza este tipo de argumentación y sostiene que una decisión a favor del individualismo lleva a asumir compromisos metafísicos además del compromiso explícito con el fisicalismo y la causación local. [Véase, Burge 1979, 1985; Loar, 1985. Fodor, 1987; Sterelny, 1990; Bilgrami, 1992, discuten la posición de Putnam y/o de Burge.]

¿Es posible superar las antinomias individualismo/no individualismo y contenido estrecho/contenido amplio? En "Más allá de la creencia" (texto del que el capítulo 10 de esta compilación es una sección), Daniel Dennett presenta como alternativa la psicología de las actitudes nocionales. Se trata de un trabajo extenso y profundo que comienza discutiendo las doctrinas a favor de las *actitudes proposicionales*. Su crítica se centra en las proposiciones y en las tres características que se les suelen atribuir: ser portadoras de valores de verdad, tener una dimensión

intensional y ser aprehendibles por la mente. Como las tres condiciones no pueden ser satisfechas de manera simultánea, la estrategia corriente ha consistido en dejar a un lado las proposiciones y adoptar una psicología de las *actitudes oracionales*. La HLP es el resultado natural de tal estrategia. Dennett critica la HLP porque 1) no permite elaborar criterios adecuados de individuación de estados psicológicos; 2) supone que uno puede descubrir la sintaxis del lenguaje del pensamiento antes de descubrir cuál es su semántica, y 3) implica que las contribuciones semánticas a un sistema siempre pueden ser presentadas en forma "verbal". Lo que se precisa, según Dennett, es una especificación de los rasgos psicológicos que sea independiente del vehículo representacional interno y del entorno externo del organismo. Esto es lo que se propone lograr mediante la psicología de las actitudes *nocionales*. El mundo nocional es el entorno para el que un organismo, siendo lo que realmente es, resulta adecuado de manera ideal. La estrategia no es nueva. Fue empleada por Husserl, es usada en la Inteligencia Artificial, está implícita en la doctrina de Quine acerca de la indeterminación de la traducción y es usada por los críticos literarios. Esta estrategia se complementa con críticas a la distinción entre actitudes proposicionales *de dicto* y *de re* [véase Simpson, 1973] y al llamado Principio de Russell, según el cual no es posible formular un juicio acerca de un objeto, sin tener conocimiento del objeto del juicio.

En "Aviso en favor de una semántica para la psicología" (capítulo 11 de esta compilación) Ned Block también se propone ofrecer un enfoque que resulte relevante para la ciencia cognitiva y que permita decidir la polémica en torno al contenido. Block adopta el funcionalismo y enumera varios *desiderata* que, según argumenta, son satisfechos por su semántica de rol conceptual. En un extenso comentario al *desideratum* 8, Block distingue el significado en sentido estrecho del significado en sentido amplio. El primero es más informativo respecto del estado mental de un sujeto. El segundo lo es respecto del sujeto en su relación con el mundo. Block evalúa las críticas de Putnam y Burge a la noción de contenido estrecho y al individualismo. También comenta el famoso enigma acerca de la creencia formulado por Kripke (el caso Pierre) [Kripke, 1979]. En definitiva, Block considera que deben reconocerse dos componentes del significado: el rol conceptual (que "está en la cabeza" e involucra el significado estrecho) y el componente externo (que abarca la relación entre representaciones/referentes y las condiciones de verdad). Se trata de una teoría de "dos factores". El trabajo culmina con una interesante clasificación de las teorías semánticas según sus impli-

caciones reduccionistas o no reduccionistas. [McGinn, 1982b, también defiende una teoría de dos factores. Véase la crítica sistemática de Bilgrami, 1992, a los intentos de "dividir" el contenido.]

Jerry Fodor vuelve sobre estos temas en "Un argumento modal en favor del contenido estrecho" (capítulo 12 de esta compilación). Fodor plantea dos argumentos que, partiendo de la premisa común de que yo y mi gemelo somos molecularmente idénticos, extraen conclusiones antinómicas respecto de la descripción intencional de nuestra conducta, de los poderes causales de nuestros estados mentales, de su pertenencia a una misma clase natural y, finalmente, respecto de la verdad del individualismo. Según Fodor, la cuestión básica es la de si los estados mentales de los gemelos moleculares pertenecen o no a clases naturales diferentes, es decir si tienen poderes causales distintos. Para dilucidar este punto, se interna en una sutil discusión acerca de la índole de las propiedades causales, los poderes causales y el carácter no conceptual del nexo. Su conclusión es que la diferencia entre los estados mentales de los gemelos no es, en realidad, una diferencia de poderes causales, ni es responsable de las diferencias que puedan darse en sus conductas. Concluye que mis estados mentales tienen que pertenecer a la misma clase natural que los de mi gemelo. En suma, el individualismo es reivindicado como verdadero.

[Para los temas tratados en las dos últimas secciones véase los trabajos compilados en Block, 1980a; Woodfield, 1982; Grimm y Merrill, 1985; Bogdan, 1986; Brand y Harnish, 1986; Silvers, 1989; Lycan, 1990. Putnam, 1989; Cummins, 1989; Fodor, 1990; Bilgrami, 1992, son algunas de las obras importantes sobre estos temas.]

7. Los modelos computacionales de la mente

La teoría del lenguaje desarrollada por Chomsky se basa en la tesis de que aplicando conjuntos de reglas recursivas a ciertas estructuras lingüísticas, se obtienen nuevas estructuras lingüísticas (adecuadas). A su vez, los programas de simulación de los procesos cognitivos realizados en computadores digitales, parten del supuesto de que la mente y el computador son sistemas de símbolos físicos y que los símbolos están codificados (físicamente) en estructuras de datos manipulados de conformidad con reglas específicas sólo sensibles a rasgos sintácticos. Ambas teorías representan a los procesos cognitivos según el "modelo de reglas y representaciones". Su influencia ha sido enorme. La unani-

midad que lograron en la comunidad científica fue tal que la base teórica de la ciencia cognitiva es prácticamente coextensiva con él. En consecuencia, los modelos filosóficos desarrollados a la vera de la ciencia cognitiva supusieron ese modelo, y una parte importante de los problemas filosóficos que se plantean en torno a la representación mental lo presuponen de manera usual. [Para una presentación clásica del modelo de reglas y representaciones, véase Pylyshyn, 1984.]

El desarrollo de modelos conexionistas ha puesto fin a ese monopolio y ha abierto una extensa polémica acerca de cuál de los dos enfoques es más adecuado para modelizar los procesos cognitivos. La "conversión" de algunos filósofos al conexionismo (Churchland, Stich, entre otros) es de por sí un hecho sintomático.

En "Una introducción al conexionismo" (capítulo 13 de esta compilación), John Tienson describe lo que da en llamar "la buena y anticuada inteligencia artificial", expone varios problemas que la afectan y describe la "crisis kuhniana" que se ha producido en su seno. En particular, Tienson analiza el problema del "marco" (determinar para un sistema de creencias, de manera efectiva y general, qué cambia con la introducción de un nuevo ítem informativo) y el problema del "cruzamiento" (reconocer que cualquier parte del sentido común tendría que ser usada en relación con cualquier tarea cognitiva). Tienson describe las características básicas de un sistema conexionista y pasa luego a una cuestión de fondo: el papel de la representación en los sistemas conexionistas. La arquitectura conexionista no reconoce la existencia de representaciones estructuradas de manera sintáctica, pero eso no es óbice, según Tienson, para que no pueda hablarse de representaciones, al menos en los ejemplos que analiza. [Sobre el conexionismo, véase McClelland y Rumelhart, 1986; Bechtel y Abrahamson, 1990; Lycan, 1990; Sterelny, 1990. Para una discusión de ambos modelos, véase Minsky y Papert, 1986.]

Fodor y Pylyshyn (1988) asumieron la defensa del modelo tradicional. Su tesis es que los modelos conexionistas del tipo PDP no permiten explicar fenómenos como la productividad de los procesos lingüísticos (la aptitud de producir y entender oraciones de estructura no limitada), la sistematicidad (entender que 'Juan es hermano de Pedro' implica 'Pedro es hermano de Juan'), la composicionalidad (el significado de una oración es una función del significado de sus partes) y la coherencia inferencial (inferir p de p & q).

En "La estructura constitutiva de los estados mentales conexionistas: una respuesta a Fodor y a Pylyshyn" (capítulo 14 de esta compila-

ción), Paul Smolensky asume la defensa del conexionismo. Define primero la Paradoja de la Cognición, es decir, el tironeo entre una concepción dura (de reglas lógicas) y una concepción blanda (de descripciones numéricas estadísticas) de la mente. Presenta cinco posiciones que pueden adoptarse frente a ella, favoreciendo la quinta opción: un sistema cognitivo es una máquina blanda y compleja en la que la dureza emerge de la blandura (el enfoque sub-simbólico). El argumento central de Fodor y Pylyshyn, según Smolensky, es que el conexionista tiene que aceptar el modelo tradicional y diseñar sus redes como meras implementaciones de éste. En respuesta, Smolensky elabora un caso de acuerdo a una propuesta crítica de Pylyshyn, y concluye que la perspectiva de la composicionalidad conexionista permite instanciar ciertos principios establecidos por los defensores del modelo clásico, "sin pasar por un lenguaje simbólico". En definitiva, la diferencia entre ambos enfoques pasaría por la distinta manera de instanciar formalmente tales principios.

En "Mentes y cerebros sin programas" (capítulo 15 de esta compilación), John Searle formula otro tipo de crítica al modelo clásico. Searle señala que la ciencia cognitiva es el candidato que hoy se considera apto para salvar el hiato entre la psicología de sentido común o psicología intencional y la neurofisiología (ha habido otros candidatos en el pasado). Critica la ecuación mente/cerebro = *software/hardware*, porque implica que no hay nada esencialmente biológico en la mente humana, y la tesis de que una máquina tiene pensamientos en el mismo sentido en que nosotros los tenemos. Su refutación se basa en "el caso de la habitación china". No puede decirse que comprendo el chino si sólo manipulo símbolos que pertenecen a ese idioma. Si tal es el caso, tampoco puede decirse que un computador digital adecuadamente programado, lo comprenda. El computador no tiene nada que yo no tenga en ese respecto. La conclusión es que instanciar el programa correcto no es suficiente para tener una mente.

En el resto del trabajo, Searle expone su tesis acerca de los fenómenos mentales. Son fenómenos causados por el cerebro y realizados en él. Para cualquier fenómeno mental hay siempre condiciones causalmente suficientes de naturaleza cerebral. Searle considera que su tesis permite lidiar con los cuatro enigmas que han atormentado, y atormentan, a los filósofos de la mente: la conciencia, la intencionalidad, la subjetividad y la causación mental. [Véase Searle, 1983, 1984, 1992. Para otras críticas de tenor parecido, véase Dreyfus, 1979.]

Hasta aquí estos extensos comentarios. No pretenden agotar los

temas tratados ni ofrecer referencias bibliográficas exhaustivas, sino servir de instrumento a quienes consideren necesario contar con un marco de referencia para los problemas de la filosofía de la mente y, en particular, para los trabajos reunidos en esta compilación.

REFERENCIAS BIBLIOGRÁFICAS

Baker, L.: (1985) "Cognitive Suicide", en Grimm y Merrill.

Baker, L.: (1987) *Saving Belief. A Critique of Physicalism*, Princeton University Press.

Beakley, B. y Ludlow, P. (comps.): (1992) *The Philosophy of Mind*, Cambridge, Mass., MIT Press.

Bechtel, W.: (1988) *Philosophy of Mind. An Overview for Cognitive Science*, Nueva York, Erlbaum.

Bechtel. W. y Abrahamson, A.: (1990) *Connectionism and the Mind. An Introduction to Parallel Processing Networks*, Oxford, Blackwell.

Bilgrami, A.: (1992) *Belief and Meaning*, Oxford, Blackwell.

Block, N y Alston, W.: (1984) "Psychology and Philosophy", en M. Bornstein (comp.), *Psychology and Its Allied Disciplines*, Nueva York, Erlbaum.

Block, N. (comp.): (1980a) *Readings in Philosophy of Psychology*, Cambridge, Mass., Harvard University Press.

Block, N.: (1980b) "What is Philosophy of Psychology?", en Block, 1980a.

Boden, M.: (1977) *Artificial Intelligence and Natural Man*, Nueva York, Basic Books.

Boden, M.: (1990) *The Philosophy of Artificial Intelligence*, Oxford, Oxford University Press.

Bogdan, R. (comp.): (1986) *Belief, Form, Content and Function*, Oxford, Clarendon.

Brand, M. y Harnish, M. (comps.): (1986) *The Representation of Knowledge and Belief*, Tucson, University of Arizona Press.

Bunge, M.: (1977) "Emergence and the Mind", *Neuroscience* 2.

Bunge, M.: (1980) *The Mind-Body Problem. A Psychobiological Approach*, Oxford, Pergamon.

Bunge, M. y Ardila, R.: (1987) *Philosophy of Psychology*, Nueva York, Springer.

Burge, T.: (1979) "Individualism and the Mental", *Midwest Studies in Philosophy* 5.

Burge, T.: (1985) "Cartesian Error and the Objectivity of Perception", en Grimm y Merrill, 1985.

Campbell, K.: (1970) *Body and Mind*, Notre Dame, Notre Dame University Press, 2da. edición, 1984.

CNRS: (1988) *Psychologie ordinaire et science cognitive*, París, CNRS.

Cummins, R.: (1989) *Meaning and Mental Representations*, Cambridge, Mass., MIT Press.

Chappell, W. (comp.): (1962) *The Philosophy of Mind*, Englewood Cliffs, Prentice-Hall.

Chomsky, N.: (1966) *Cartesian Linguistics*, Nueva York, Harper & Row.

Chomsky, N.: (1979) *Reflections on Language*, Nueva York, Pantheon.

Churchland, P.M.: (1979) *Scientific Realism and the Plasticity of Mind*, Cambridge, Cambridge University Press.

Churchland, P.M.: (1988) *Matter and Consciousness*, Cambridge, Mass., MIT Press.

Churchland, P.M.: (1991) "Folk Psychology and the Explanation of Human Behavior", en Greenwood, 1991.

Churchland, P.S.: (1986) *Neurophilosophy. Toward a Unified Theory of Mind and Brain*, Cambridge, Mass., MIT Press.

Dennett, D.: (1978) *Brainstorms*, Montgomery, Bradford Books.

Dennett, D.: (1987) *The Intentional Stance*, Cambridge, Mass., MIT Press.

Dennett, D.: (1991) "Two Contrasts: Folk Crafts vs. Folk Science and Belief vs. Opinion", en Greenwood, 1991.

Dretske, F.: (1988) *Explaining Behavior. Reasons in a World of Causes*, Cambridge, Mass., MIT Press.

Dreyfus, H.: (1979) *What Computers Can't Do*, Nueva York, Harper & Row.

Eckardt, B. von: (1984) "Cognitive Psychology and Principled Skepticism", *Journal of Philosophy*, 81.

Feyerabend, P.: (1963) "Mental Events and the Brain", *Journal of Philosophy*, 60.

Field, H.: (1978) "Mental Representations", *Erkenntnis* 13.

Flanagan, O.: (1984) *The Science of the Mind*, Cambridge, Mass., MIT Press.

Fodor, J.: (1975) *The Language of Thought*, Nueva York, Crowell.

Fodor, J.: (1980) "Methodological Solipsism Considered as a Research Program in Cognitive Psychology", *Behavioral and Brain Sciences* 3.

Fodor, J.: (1981) *Representations*, Cambridge, Mass., MIT Press.
Fodor, J.: (1985) "Fodor's Guide to Mental Representation", *Mind* 9.
Fodor, J.: (1987) *Psychosemantics*, Cambridge, Mass., MIT Press.
Fodor, J. y Pylyshyn, Z.: (1988) "Connectionism and Cognitive Arquitecture. A Critical Analysis", *Cognition* 28.
Fodor, J.: (1990) *A Theory of Content and Other Essays*, Cambridge, Mass., MIT Press.
Gardner, H.: (1985) *The Mind's New Science*, Nueva York, Basic Books.
Globus, G., Maxwell, G. y Sadovnik, I. (comps.): (1976) *Consciousness and the Brain*, Nueva York, Plenum.
Greenwood, J. (comp.): (1991) *The Future of Folk Psychology. Intentionality and Cognitive Science*, Cambridge, Cambridge University Press.
Grimm, R. y Merrill, D. (comps.): (1985) *Contents of Thougths*, Tucson, University of Arizona Press.
Guftafson, D. (comp.): (1964) *Philosophical Psychology*, Garden City, Dubleday.
Hampshire, S. (comp.): (1966) *Philosophy of Mind*, Nueva York, Harper & Row.
Harman, G.: (1975) *Thought*, Princeton, Princeton University Press.
Haugeland, J. (comp.): (1981) *Mind Design. Philosophy, Psychology and Artificial Intelligence*, Cambridge, Mass., MIT Press.
Haugeland, J.: (1985) *Artificial Intelligence. The Very Idea*, Cambridge, Mass., MIT Press.
Horgan, T. y Woodward, J.: (1985) "Folk Psychology Is Here to Stay", *Philosophical Review* 94. Incluido en Greenwood, 1991.
Jackson, F. y Petit, P.: (1990) "In Defense of Folk Psychology", *Philosophical Studies* 59.
Kitcher, P.: (1984) "In Defense of Intentional Psychology", *Journal of Philosophy* 81.
Kripke, S.: (1979) "A Puzzle about Belief", en Margalit 1979.
Kripke, S.: (1980) *Naming and Necessity*, Cambridge, Mass., Harvard University Press.
Lindsay, N. y Norman, D.: (1983) *Human Information Processing. An Introduction to Psychology*, Nueva York, Academic Press.
Loar, B.: (1985) "Social Content and Psychological Content", en Grimm y Merrill, 1985.
Lycan, W.: (1988) *Judgment and Justification*, Cambridge, Cambridge University Press.
Lycan, W. (comp.): (1990) *Mind and Cognition*, Oxford, Blackwell.

Margalit, A. (comp.): (1979) *Meaning and Use*, Dordrecht, Reidel.
McClelland, J., Rumelhart, D. y PDP Group: (1986) *Parallel Distributed Processing. Exploration in the Microstructure of Cognition*, Cambridge, Mass., MIT Press.
McDonough, R.: (1991) "A Culturalist Account of Folk Psychology", en Greenwood, 1991.
McGinn, C.: (1982a) *The Character of Mind*, Oxford, Oxford University Press.
McGinn, C.: (1982b) "The Structure of Content", en Woodfield, 1982.
Meyering, T.: (1989) *The Historical Roots of Cognitive Science*, Dordrecht, Kluger.
Millikan, R.: (1984) *Language, Thought and Other Biological Categories*, Cambridge, Mass., MIT Press.
Millikan, R.: (1986) "Thoughts without Laws. Cognitive Science without Content", *Philosophical Review* 95.
Millikan, R.: (1989) "In Defense of Proper Functions", *Philosophy of Science* 56.
Minsky, M. y Papert, S.: (1986) *Perceptrons*, Cambridge, Mass., MIT Press., ed. ampliada.
Nudler, O. (comp.): (1975) *Problemas Epistemológicos de la Psicología*, Buenos Aires, Siglo XXI.
Pérez, D.: (1993) "Sentido común y psicología. Notas sobre la psicología de sentido común", *Cuadernos de Filosofía* 13.
Putnam, H.: (1975a) *Mind, Language and Reality. Philosophical Papers II*, Londres, Cambridge University Press.
Putnam, H.: (1975b) "The Meaning of 'Meaning' ", en Putnam, 1975a.
Putnam, H.: (1989) *Representation and Reality*, Cambridge, Mass., MIT Press.
Pylyshyn, Z.: (1984) *Computation and Cognition. Toward a Foundation of Cognitive Science*, Cambridge, Mass., MIT Press.
Rabossi, E.: (1979) "¿Por qué el sentido común importa a la filosofía?", *Manuscrito* 3.
Ramsay, W., Stich S. y Garon, J.: (1991) "Connectionism, Eliminativism and the Future of Folk Psychology", en Greenwood, 1991.
Ryle, G.: (1949) *The Concept of Mind*, Londres, Hutchinson.
Schiffer, S.: (1987) *Remnants of Meaning*, Cambridge, Mass., MIT Press.
Searle, J.: (1983) *Intentionality. An Essay in the Philosophy of Mind*, Cambridge, Cambridge University Press.
Searle, J.: (1984) *Minds, Brains and Science*, Cambridge, Mass., Harvard University Press.

Searle, J.: (1992) *The Rediscovery of the Mind*, Cambridge, Mass., MIT Press.

Sellars, W.: (1958) "Empiricism and the Philosophy of Mind", en Sellars, 1963.

Sellars, W.: (1963) *Science, Perception and Reality*, Londres, Routledge.

Silvers, S. (comp.): (1989) *Representations. Readings in the Philosophy of Mental Representation*, Dordrecht, Reidel.

Simon, H.: (1969) *Sciences of the Artificial*, Cambridge, Mass., MIT Press.

Simpson, T.: (1973) *Semántica filosófica. Problemas y discusiones*, Buenos Aires, Siglo XXI.

Sober, E.: (1985) "Putting the Functions Back in Functionalism", *Synthèse* 64.

Sterelny, K.: (1990) *The Representational Theory of Mind*, Cambridge, Mass., Blackwell.

Stich, S.: (1978) "Autonomous Psychology and the Belief-Desire Thesis", *Monist* 61.

Stich, S.: (1982) "On the Adscription of Content", en Woodfield, 1982.

Stich, S.: (1983) *From Psychology to Cognitive Science*, Cambridge, Mass., MIT Press.

Stillings, N. *et al.*: (1987) *Cognitive Science*, Cambridge, Mass., MIT Press.

Vega, M. de: (1987) *Introducción a la Psicología Cognitiva*, Madrid, Alianza.

Wimsatt, W.: (1976) "Reductionism, Level of Organization and the-Mind-Body Problem", en Globus *et al.*, 1976.

Woodfield, A. (comp.): (1982) *Thought and Object*, Oxford, Clarendon.

II

EL *STATUS* DE LA PSICOLOGÍA DE SENTIDO COMÚN *VIS-À-VIS*, LA PSICOLOGÍA COGNITIVA Y LAS NEUROCIENCIAS

Capítulo 2

EL MATERIALISMO ELIMINATIVO Y LAS ACTITUDES PROPOSICIONALES *

Paul M. Churchland **

El materialismo eliminativo es la tesis que sostiene que nuestra concepción de sentido común acerca de los fenómenos psicológicos constituye una teoría radicalmente falsa, una teoría tan esencialmente defectuosa que tanto sus principios como su ontología serán eventualmente desplazados, más que reducidos con fluidez, por una neurociencia completa [*completed neuroscience*]. Nuestra comprensión mutua y aun nuestra introspección podrán ser entonces reconstituidas dentro del marco conceptual de la neurociencia completa; una teoría que esperamos sea mucho más poderosa que la psicología de sentido común [*common-sense psychology*] a la que desplaza y, en general, mucho más integrada a la ciencia física. Mi propósito en este artículo es explorar esas proyecciones, especialmente las que atañen a 1) los elementos principales de la psicología de sentido común: las actitudes proposicionales (creencias, deseos, etcétera), y 2) la concepción de la racionalidad en la que esos elementos figuran.

Este enfoque representa un cambio en la suerte del materialismo. Hace veinte años se consideraba que las emociones, los *qualia* y las "vivencias puras" ["*raw feels*"] eran los obstáculos principales para el programa materialista. Al disolverse[1] estas barreras, el *locus* de la opo-

* "Eliminative Materialism and Propositional Attitudes", *The Journal of Philosophy* 78 (1981), págs. 67-90. Con autorización del autor y del *Journal of Philosophy*.

** Una versión anterior de este artículo fue presentada en la Universidad de Ottawa y en el coloquio *Brain, Mind and Person* en SUNY, Oswego. Agradezco las sugerencias y críticas que han dado forma a la presente versión.

1. Véase Paul Feyerabend, "Materialism and the Mind-Body Problem", *Review of Metaphysics*, XVII.1, 65 (septiembre de 1963): 49-66; Richard Rorty, "Mind-Body Identity, Privacy and Categories", *ibid.*, XIX.1, 73 (septiembre de 1965), 24-54 y mi *Scientific Realism and the Plasticity of Mind* (Nueva York: Cambridge, 1979).

sición ha cambiado. Ahora se sostiene que es el ámbito de lo intencional, el ámbito de la actitud proposicional, el que es irreductible a todo lo que pertenezca a un marco materialista, e ineliminable respecto de él. Tenemos que examinar si esto es así, y por qué lo es.

Tal examen no tendría mucho sentido, sin embargo, a menos que previamente se reconozca que la red pertinente de los conceptos del sentido común constituye realmente una teoría empírica, con todas las funciones, las virtudes *y los peligros* que implica esta condición. Por lo tanto, comenzaré con un breve esbozo de ese punto de vista y con una enumeración sucinta de sus fundamentos. Me sorprende la resistencia que aún encuentra. Después de todo el sentido común ha producido muchas teorías. Recordemos el punto de vista según el cual el espacio tiene una dirección natural para todas las cosas que caen; que el peso es una propiedad intrínseca de los cuerpos; que un móvil no sometido a fuerzas rápidamente volverá al estado de reposo; que la esfera de los cielos gira a diario, etcétera. Estos ejemplos son claros, quizá, pero la gente parece dispuesta a conceder un componente teórico al sentido común sólo si 1) la teoría y el sentido común involucrados pueden ser ubicados inocuamente en la antigüedad, y 2) la teoría pertinente resulta ahora falsa, con tanta claridad, que su naturaleza especulativa es ineludible. Las teorías son, por cierto, más fáciles de advertir en esas circunstancias. Pero la visión de lo ya ocurrido es siempre perfecta. Para cambiar un poco, aspiremos a formular ciertas premoniciones.

1. Por qué la psicología folk *es una teoría*

Considerar a nuestro marco conceptual de sentido común para los fenómenos mentales como una teoría, proporciona una organización simple y unificadora de la mayoría de los tópicos principales de la filosofía de la mente, que incluye la explicación y la predicción de la conducta, la semántica de los predicados mentales, la teoría de la acción, el problema de las otras mentes, la intencionalidad de los estados mentales, la naturaleza de la introspección y el problema mente-cuerpo. Cualquier punto de vista que pueda abarcar todo esto merece una consideración cuidadosa.

Comencemos con la explicación de la conducta humana (y animal). El hecho es que las personas comunes son capaces de explicar e incluso de predecir con notable facilidad y éxito la conducta de otras personas. Por lo común, tales explicaciones y predicciones hacen referencia a los

deseos, creencias, temores, intenciones, percepciones, etc, que se supone que los agentes padecen. Pero las explicaciones presuponen leyes —leyes toscas y operativas, al menos— que conecten las condiciones explicativas con la conducta explicada. Lo mismo vale para la formulación de predicciones y para la justificación de los condicionales subjuntivos y contrafácticos que conciernen a la conducta. Afortunadamente, puede reconstruirse una rica red de leyes de sentido común a partir de este intercambio cotidiano de explicación y anticipación; sus principios son preceptos familiares y sus variadas funciones son transparentes. Cada uno de nosotros comprende tan bien a los demás porque compartimos el dominio tácito de un cuerpo integrado de saberes populares [*lore*] que conciernen a las relaciones legaliformes que valen entre circunstancias externas, estados internos y conducta pública. Dada su naturaleza y funciones, ese cuerpo de saberes puede ser llamado, acertadamente, "psicología *folk*" [*folk psychology*].[2]

Este enfoque implica que la semántica de los términos de nuestro vocabulario familiar mentalista debe entenderse, en general, del mismo modo que la semántica de los términos teóricos: el significado de cualquier término teórico está fijado o constituido por la red de leyes en la que figura. (Esta posición es muy distinta del conductismo lógico. Negamos que las leyes relevantes sean analíticas; en general, son las conexiones legaliformes las que llevan el peso semántico y no sólo las conexiones con la conducta pública. Esta perspectiva da cuenta del mínimo de plausibilidad de que gozó el conductismo lógico.)

Lo que es más importante aún, el reconocimiento de que la psicología *folk* es una teoría proporciona una solución simple y decisiva a un viejo problema escéptico: el problema de las otras mentes. La convicción problemática de que otro individuo sea el sujeto de ciertos estados mentales no se infiere deductivamente de su conducta, ni se infiere por analogía inductiva a partir del ejemplo, peligrosamente aislado, del caso propio. Más bien, esa convicción es una *hipótesis explicativa* singular de un tipo perfectamente simple. Su función, en conjunción con las leyes de fondo de la psicología *folk* es la de proporcionar explicaciones/ predicciones/ comprensión de la conducta continua de los individuos, y es creíble en la medida en que resulte más exitosa, en ese respecto, que

2. En seguida examinaremos un manojo de estas leyes. Para una ejemplificación más abarcadora de las leyes de la psicología de sentido común, véase mi *Scientific Realism and the Plasticity of Mind, op.cit.*, cap. 4. Para un examen detallado de los principios de la PSC que suscriben explicaciones de acciones en particular, véase mi "The Logical Character of Action Explanations", *Philosophical Review*, LXXIX, 2 (abril de 1970): 214-236.

otras hipótesis alternativas. En lo fundamental, tales hipótesis son exitosas y por lo tanto la creencia de que los demás gozan de los estados internos concebidos por la psicología *folk*, resulta ser razonable.

Así, el conocimiento de las otras mentes no tiene una dependencia esencial del conocimiento de nuestra propia mente. Al aplicar los principios de nuestra psicología *folk*, un marciano podría adscribirnos, debidamente, la serie familiar de estados mentales, aun cuando su propia psicología fuera muy diferente de la nuestra. En consecuencia él no estaría "generalizando a partir del caso propio".

Del mismo modo, los juicios introspectivos acerca del caso propio resultan no tener ningún *status* o integridad especial. Desde esta perspectiva, un juicio introspectivo es sólo un ejemplo de un hábito adquirido de respuesta conceptual a los estados internos propios, y la integridad [*integrity*] de cualquier respuesta particular siempre es contingente con respecto a la integridad del marco conceptual adquirido (teoría) en el que la respuesta está enmarcada. Por consiguiente, la certidumbre *introspectiva* que uno tiene de que la mente propia es el asiento de creencias y deseos, puede estar tan fuera de lugar como lo estuvo la certidumbre *visual* del hombre clásico según la cual la esfera de los cielos, salpicada de estrellas, se mueve a diario.

Otro problema es la intencionalidad de los estados mentales. Las 'actitudes proposicionales', como Russell las llamó, forman el núcleo sistemático de la psicología *folk*. Su singularidad y sus propiedades lógicas anómalas han llevado a algunos a advertir un contraste fundamental con todo lo que los meros fenómenos físicos podrían concebiblemente exhibir. La clave de este asunto consiste nuevamente en la naturaleza teórica de la psicología *folk*. La intencionalidad de los estados mentales emerge aquí no como un misterio de la naturaleza sino como un rasgo estructural de los conceptos de la psicología *folk*. Irónicamente esos mismos rasgos estructurales revelan la estrecha afinidad que la psicología *folk* guarda con las teorías de las ciencias físicas. Permítaseme intentar explicar esto.

Considérese la gran variedad de lo que se puede llamar "actitudes numéricas", que aparecen en el marco conceptual de las ciencias físicas: '... tiene una masa $_{kg}$ de *n*', '...tiene una velocidad de *n*', '...tiene una temperatura $_{k}$ de *n*', etcétera. Estas son expresiones formadoras-de-predicados [*predicate forming expressions*]: cuando uno ubica un término singular numérico [*singular term for a number*] en el lugar ocupado por '*n*', resulta un predicado determinado. Lo que es más interesante, las relaciones que se dan entre las diversas "actitudes numéricas", son pre-

cisamente las relaciones que se dan entre los números "contenidos" en esas actitudes. Lo que es aún más interesante, el lugar de argumento [*argument place*] que ocupan los términos singulares numéricos es susceptible de cuantificación. Todo esto permite la expresión de generalizaciones concernientes a las relaciones legaliformes que valen en la naturaleza entre las diversas actitudes numéricas. Tales leyes involucran la cuantificación sobre números y aprovechan las relaciones matemáticas que valen en ese dominio. Así por ejemplo,

(1) (x) (f) (m) (((x tiene una masa de m) & (x sufre una fuerza neta de f)) ⊃ (x se acelera a f/m)).

Considérese ahora la gran variedad de actitudes proposicionales: '...cree que p', '...desea que p', '...teme que p', '...está contento de que p', etcétera. Estas expresiones son también expresiones formadoras-de-predicados. Cuando se ubica un término singular para proposiciones en el lugar ocupado por 'p', resulta un predicado determinado, por ejemplo '...cree que Juan es alto'. (Las oraciones no funcionan generalmente como términos singulares, pero es difícil escapar a la idea de que cuando aparece una oración en el lugar ocupado por 'p', funcione como un término singular o como si lo fuera. Sobre esto, ver más adelante.) Lo que es más interesante, las relaciones entre las actitudes proposicionales resultantes son, característicamente, las relaciones que valen entre las proposiciones "contenidas" en ellas, relaciones tales como la implicación, la equivalencia y la inconsistencia. Lo que es aún más interesante, el lugar de argumento que toma proposiciones como términos singulares es susceptible de cuantificación. Todo esto permite la expresión de generalizaciones que conciernen a las relaciones legaliformes que valen entre las actitudes proposicionales. Tales leyes involucran la cuantificación sobre proposiciones y aprovechan varias relaciones que valen en ese dominio. Así por ejemplo:

(2) (x) (p) ((x teme que p) ⊃ (x desea que —p))
(3) (x) (p) ((x espera que p) & (x descubre que p) ⊃ (x se complace de que p))
(4) (x) (p) (q) ((x cree que p) & (x cree que (si p entonces q))) ⊃ (salvo confusión, distracción, etcétera, x cree que q))

(5) $(x)\ (p)\ (q)$ (((*x* desea que *p*) & (*x* cree que (si *q* entonces *p*)) & (*x* es capaz de producir *q*) ⊃ (salvo deseos antagónicos o estrategias preferibles, *x* produce *q*)).[3]

La psicología *folk* no sólo es una teoría, sino que lo es de una manera tan *obvia* que resulta un gran misterio por qué los filósofos lo han advertido recién en la última mitad del siglo veinte. Los rasgos estructurales de la psicología *folk* son perfectamente paralelos a los de la física matemática; la única diferencia reside en el dominio de las entidades abstractas que cada una aprovecha: números en el caso de la física y proposiciones en el caso de la psicología.

Finalmente, advertir que la psicología *folk* es una teoría echa nueva luz sobre el problema mente-cuerpo. El problema deviene entonces en la cuestión de cómo la ontología de una teoría (la psicología *folk*) estará o no estará relacionada con la ontología de otra teoría (la neurociencia completa), y las principales posiciones filosóficas sobre el problema mente-cuerpo emergen así como diferentes anticipaciones de lo que la investigación futura revelará acerca del *status* interteórico y de la corrección [*integrity*] de la psicología *folk*.

El teórico de la identidad espera, con optimismo, que la psicología *folk* sea fácilmente *reducida* a la neurociencia completa y que su ontología se preserve en virtud de identidades transteóricas. El dualista espera que resulte *irreducible* a la neurociencia completa, en virtud de una descripción no redundante de un dominio no físico, autónomo, de los fenómenos naturales. El funcionalista también espera que resulte irreducible, pero sobre la base de fundamentos muy diferentes: la economía interna caracterizada por la psicología *folk* no es, en último análisis, una

3. Si se permanece dentro de una interpretación objetal de los cuantificadores, quizá la manera más simple de dar sentido sistemático a expresiones como '*x* cree que *p*' y oraciones cerradas formadas a partir de ellas, consiste exactamente en interpretar todo lo que acaece en la posición encerrada [*nested*] ocupada por '*p* , *q*' etcétera, como si tuvieran la función de un término singular. Por esta razón los conectivos habituales, tal como aparecen entre términos en esa posición encerrada, deben ser interpretados como si estuvieran funcionando como operadores que forman términos singulares compuestos a partir de otros términos singulares y no como operadores oracionales. Los términos singulares compuestos así formados denotan las proposiciones compuestas adecuadas. La cuantificación sustitucional, por supuesto respaldará una interpretación diferente, y también hay otros enfoques. Especialmente atractivo es el enfoque pro-oracional de Dorothy Grover, Joseph Camp y Nuel Belnap, "A Prosentential Theory of Truth", *Philosophical Studies*, XXVII, 2 (febrero de 1975): 73-125. Pero la resolución de estos temas no es vital para la presente discusión.

economía de estados naturales gobernados por leyes, sino una organización abstracta de estados funcionales, una organización instanciable [*instantiable*] en una variedad de sustratos materiales muy diferentes. Y por tanto, irreducible a los principios peculiares de cualquiera de ellos.

Finalmente, el materialista eliminativo es también pesimista acerca de las perspectivas de reducción, pero su razón es que la psicología *folk* es una explicación radicalmente inadecuada de nuestras actividades internas, demasiado confusa y demasiado defectuosa como para sobrevivir a una reducción interteórica. Desde su punto de vista, será desplazada, simplemente, por una teoría mejor de tales actividades.

Cuál de estas suertes será el verdadero destino de la psicología *folk*, es lo que trataremos de pronosticar. Por ahora, el punto a tener presente es que vamos a explorar la suerte de una teoría, una *teoría* especulativa, corregible y sistemática.

2. *Por qué la psicología* folk *podría (realmente) ser falsa*

Dado que la psicología *folk* es una teoría empírica, existe al menos la posibilidad abstracta de que sus principios sean radicalmente falsos y de que su ontología sea una ilusión. Sin embargo, con la excepción del materialismo eliminativo, ninguna de las principales posiciones considera seriamente tal posibilidad. Ninguna de ellas duda de la integridad básica o de la verdad de la psicología *folk* (de aquí en más, PF), y todas ellas prevén un futuro en el que se conservarán sus leyes y categorías. Este conservadurismo no carece de algún fundamento. Después de todo la PF disfruta de un éxito explicativo y predictivo substancial. ¿Y qué mejor fundamento para confiar en la integridad de sus categorías?

Realmente, ¿qué mejor fundamento? Aún así, la presunción a favor de la PF es espuria, nace de una visión inocente y estrecha. Un examen más cuidadoso revela una imagen diferente. Primero, debemos considerar a la PF no sólo en relación con sus éxitos sino también con sus defectos explicativos, sus alcances y su seriedad. Segundo, debemos considerar la prolongada historia de la PF, su crecimiento, fertilidad y promesa efectiva de desarrollo futuro. Y tercero, debemos considerar qué tipos de teorías de la etiología de nuestra conducta son *probablemente* verdaderas, dado todo lo que hemos aprendido en épocas recientes acerca de nosotros mismos. Esto es, debemos evaluar a la PF respecto de su continuidad y coherencia con teorías fértiles y bien establecidas en dominios adyacentes y superpuestos —por ejemplo, con la teoría de la

evolución, la biología y la neurociencia—, porque la coherencia activa con el resto de lo que presumimos conocer es, quizá, la medida última de cualquier hipótesis.

Un inventario serio de este tipo revela una situación muy problemática que concitaría un neto escepticismo en el caso de cualquier teoría menos familiar y menos apreciada por nosotros. Permítasenos esquematizar algunos detalles pertinentes. Cuando centramos nuestra atención, no sobre lo que la PF puede explicar, sino sobre lo que no puede explicar o incluso pasa por alto, descubrimos que hay mucho para elaborar. Como ejemplos de fenómenos mentales importantes y centrales que quedan en el misterio, total o parcialmente, en el marco de la PF, consideremos la naturaleza y la dinámica de la enfermedad mental, la facultad de la imaginación creadora o el fundamento de las diferencias de inteligencia entre los individuos. Consideremos nuestra total ignorancia de la naturaleza y las funciones psicológicas del sueño, ese curioso estado en el que pasamos un tercio de nuestra vida. Reflexionemos sobre la habilidad común de atajar al vuelo una pelota mientras corremos, o acertar a un auto en movimiento con una bola de nieve. Consideremos la construcción interna de una imagen visual tridimensional en nuestras respectivas retinas, a partir de diferencias sutiles en las organizaciones bidimensionales de los estímulos. Consideremos la rica variedad de ilusiones perceptuales, visuales y de otro tipo. O consideremos el milagro de la memoria, con su centelleante capacidad para recuperar lo pertinente. La PF ilumina muy poco estos y muchos otros fenómenos mentales.

Un misterio particularmente destacable es el de la naturaleza del proceso mismo de aprendizaje, especialmente cuando involucra un cambio conceptual en gran escala y cuando aparece en su forma prelingüística o en su forma no-lingüística (como en los bebés o en los animales) que es, con mucho, la forma más común en la naturaleza. La PF se enfrenta en estos casos con dificultades especiales puesto que su concepción del aprendizaje, como la manipulación y el almacenamiento de actitudes proposicionales, fracasa ante el hecho de que formular, manipular y almacenar una rica trama de actitudes proposicionales es en sí mismo algo que se aprende, y es sólo una entre muchas otras habilidades cognitivas adquiridas. La PF aparecería así como constitucionalmente incapaz de abordar siquiera este misterio sumamente básico.[4]

4. Una respuesta posible aquí es insistir en que la actividad cognitiva de los animales y los bebés es linguaforme en sus elementos, estructuras y procesamiento, ya desde el

Fracasos de tal magnitud no muestran (todavía) que la PF sea una teoría falsa, pero sí elevan esa probabilidad al rango de una posibilidad real y muestran decisivamente que la PF es, *en el mejor de los casos*, una teoría altamente superficial, una glosa parcial y no penetrante acerca de una realidad más compleja y más profunda. Habiendo alcanzado esta conclusión se nos puede disculpar que exploremos la posibilidad de que la PF proporcione un esquema realmente engañoso de nuestra cinemática [*kinematics*] y dinámica internas, cuyo éxito se debe más a una aplicación selectiva y a una interpretación forzada de nuestra parte, que a una comprensión teórica genuina por parte de la PF.

Una mirada a la historia de la PF aquieta poco tales temores, una vez que han surgido. Es una historia de retraimiento, infertilidad y decadencia. El dominio presunto de la PF solía ser mucho mayor que el que es ahora. En las culturas primitivas, la conducta de la mayoría de los elementos de la naturaleza era comprendida en términos intencionales. El viento podía enojarse, la luna ser celosa, el río ser generoso, el mar enfurecerse, etcétera. Éstas no eran metáforas. Se hacían sacrificios y se formulaban augurios para aplacar o adivinar las pasiones cambiantes de los dioses. A pesar de su esterilidad, este enfoque animista de la naturaleza ha dominado nuestra historia y sólo en los últimos dos o tres mil años hemos reducido la aplicación literal de la PF al dominio de los animales superiores.

Sin embargo, incluso en este dominio preferencial, tanto el contenido como el éxito de la PF no han avanzado perceptiblemente en dos o tres mil años. La PF de los griegos es esencialmente la PF que usamos hoy, y es mínima la diferencia entre nuestra capacidad de explicar la conducta humana y la que tenía Sófocles. Para cualquier teoría, éste es un período muy largo de estancamiento e infertilidad, especialmente cuando se enfrenta a tamaña lista de anomalías y misterios en su propio dominio explicativo. Quizá las teorías perfectas no tengan que evolucionar. Pero la PF es profundamente imperfecta. Su incapacidad para desarrollar sus recursos y extender su ámbito de logros es, en consecuencia, sospechosamente curiosa, y uno tiene que cuestionar la integridad de sus categorías básicas. Para usar los términos de Imre Lakatos, la PF es

---

nacimiento. J. A. Fodor, en *The Language of Thought* (Nueva York, Crowell, 1975) ha propuesto una teoría positiva del pensamiento bajo el supuesto de que las formas innatas de la actividad cognitiva tienen precisamente la forma que aquí se les niega. Para una crítica de la postura de Fodor, véase Patricia Churchland, "Fodor on Language Learning", *Synthèse*, XXXVIII, 1 (mayo de 1978), 149-159.

un programa de investigación degenerado o estancado, y lo ha sido durante milenios.

El éxito explicativo efectivo no es, por supuesto, la única dimensión en la que una teoría puede ser virtuosa o prometedora. Una teoría estancada o problemática puede merecer paciencia y atención debido a otros fundamentos; por ejemplo, debido a que es la única teoría o enfoque teórico que se ajusta bien a otras teorías acerca de contenidos adyacentes, o la única que promete reducirse a o ser explicada por alguna teoría de fondo ya establecida, cuyo dominio incluya el dominio de la teoría en cuestión. En suma, puede ser digna de crédito porque promete una integración teórica. ¿Cómo valorar a la PF en esta dimensión?

Quizá sea precisamente aquí donde la PF muestra su desempeño más bajo. Si encaramos al *homo sapiens* desde la perspectiva de la historia natural y de las ciencias físicas, podemos ofrecer un relato [*story*] coherente de su constitución, desarrollo y capacidades conductuales que abarque la física de partículas, la teoría atómica y molecular, la química orgánica, la teoría de la evolución, la biología, la fisiología y la neurociencia materialista. Ese relato, aunque todavía radicalmente incompleto, es ya extremadamente poderoso y superador de la PF en muchos puntos, aun en su propio dominio. Y es coherente, deliberada y autoconscientemente, con el resto de nuestra imagen del mundo, en constante desarrollo. Con pocas palabras, la mayor síntesis teórica en la historia de la raza humana se encuentra ya en nuestras manos, y partes de ella nos proporcionan descripciones y explicaciones agudas del *input* sensorial, la actividad neural y el control motor humanos.

Pero la PF no forma parte de esta síntesis creciente. Sus categorías intencionales permanecen grandilocuentemente aisladas, sin perspectiva visible de reducción a ese *corpus* más amplio. Desde mi punto de vista, una reducción exitosa no puede ser descartada, pero la impotencia explicativa de la PF y su largo estancamiento no dan pie para esperar que sus categorías vayan a verse prolijamente reflejadas en el marco de la neurociencia. Por el contrario, nos recuerda cómo la alquimia tiene que haber sido vista cuando la química elemental iba tomando forma, cómo la cosmología aristotélica tiene que haber sido vista cuando se iba articulando la mecánica clásica, o cómo la concepción vitalista de la vida tiene que haber sido vista cuando progresaba la química orgánica.

Al bosquejar un resumen apropiado de esta situación, debemos hacer un esfuerzo especial para hacer abstracción del hecho de que la PF es una parte central de nuestra *Lebenswelt* habitual y de que sirve como el vehículo principal de nuestros intercambios interpersonales.

Porque esos hechos dan a la PF una inercia conceptual que excede sus virtudes puramente teóricas. Si nos limitamos a esta última dimensión, lo que tenemos que decir es que la PF padece fracasos explicativos en una escala épica, que ha estado estancada durante por lo menos veinticinco siglos y que sus categorías parecen (hasta ahora) ser inconmensurables u ortogonales respecto de las categorías de la ciencia física de fondo, cuya antigua pretensión de explicar la conducta humana parece innegable. Debe aceptarse que cualquier teoría que se ajuste a esta descripción es un candidato serio para una franca eliminación.

Por supuesto que en esta etapa no podemos insistir en una conclusión más fuerte; ni es mi interés hacerlo. Aquí estamos explorando una posibilidad y los hechos exigen que se la tome en serio; ni más ni menos. El rasgo distintivo del materialista eliminativo es considerarla muy seriamente.

3. *Los argumentos en contra de la eliminación*

El fundamento básico del materialismo eliminativo es éste: la PF es una teoría, y muy probablemente una teoría falsa; intentemos, por lo tanto, ir más allá de ella.

El fundamento es claro y simple, pero muchos no lo encuentran convincente. Se objetará que la PF estrictamente hablando no es una teoría *empírica*, que no es falsa o que por lo menos no es refutable por consideraciones empíricas, y que no debe o que no puede ser trascendida a la manera de una teoría empírica difunta. En lo que sigue examinaremos estas objeciones tal como surgen de la más popular y mejor fundada de las posiciones que compiten en la filosofía de la mente: el funcionalismo.

Cierta aversión al materialismo eliminativo emana de dos líneas distintas que se desarrollan en el funcionalismo contemporáneo. La primera concierne al carácter *normativo* de la PF, o al menos a ese núcleo central de la PF que trata de las actitudes proposicionales. Algunos dirán que la PF es la caracterización de un ideal o, al menos, de un modo plausible de actividad interna. Delinea no sólo lo que es tener y procesar creencias y deseos, sino también (e inevitablemente), lo que es ser racional en el gobierno de ellos. El ideal establecido por la PF puede ser alcanzado imperfectamente por los seres humanos empíricos, pero esto no impugna a la PF en tanto que caracterización normativa. Ni es necesario que tales fracasos impugnen seriamente a la PF aun como una caracte-

rización descriptiva, ya que sigue siendo verdadero que nuestras actividades pueden ser comprendidas como racionales, de manera provechosa y precisa, *salvo por* lapsos ocasionales debidos a ruidos, interferencias u otras fallas; defectos que la investigación empírica puede eventualmente aclarar. Por tal razón, aunque la neurociencia pueda enriquecerla provechosamente, la PF no tiene una necesidad imperiosa de ser desplazada aún como teoría descriptiva; ni podría ser reemplazada *qua* caracterización normativa por ninguna teoría descriptiva de los mecanismos neurales, puesto que la racionanalidad se define sobre la base de actitudes proposicionales tales como creencias y deseos. Por lo tanto, la PF seguirá estando con nosotros [*is here to stay*].

Daniel Dennett ha defendido una posición similar.[5] Además, el punto de vista que se acaba de delinear se hace eco también de un tema de los dualistas de propiedades. Karl Popper y Joseph Margolis señalan a la naturaleza normativa de la actividad mental y lingüística como un impedimento para intentar su penetración o eliminación mediante una teoría materialista o descriptiva.[6] Más adelante espero desacreditar la atracción de tales propuestas.

La segunda línea concierne a la naturaleza *abstracta* de la PF. La pretensión central del funcionalismo es que los principios de la PF caracterizan nuestros estados internos de un modo tal que no hacen referencia a su naturaleza intrínseca o a su constitución física. Son caracterizados, en cambio, en términos de la red de relaciones causales que mantienen entre sí y con las circunstancias sensoriales y la conducta pública. Dada su especificación abstracta, esa economía interna puede en consecuencia ser realizada en una variedad nómicamente heterogénea de sistemas físicos. Todos ellos pueden diferir, aun radicalmente, en su constitución física y, sin embargo, en otro nivel, todos ellos compartirán la misma naturaleza. Este punto de vista, dice Fodor, "es compatible con alegaciones muy fuertes acerca de la ineliminabilidad del lenguaje mental en las teorías conductistas".[7] Dada la posibilidad efectiva de instanciaciones múltiples en sustratos físicos heterogéneos, no pode-

5. Muy explícitamente en "Three Kinds of Intentional Psychology (cap. 3 de *The Intentional Stance*, Cambridge, Mass., MIT, 1987), pero este tema de Dennett se encuentra ya en sus "Intentional Systems", *The Journal of Philosophy*, LXVIII, 4 (feb. 25, 1971), 87-106; reimpreso en su *Brainstorms* (Montgomery, Vt., Bradford Books, 1978).

6. Popper, *Objective Knowledge* (Nueva York, Oxford, 1972); con J. Eccles, *The Self and Its Brain* (Nueva York, Springer Verlag, 1978). Margolis, *Persons and Minds* (Boston, Reidel, 1978).

7. *Psychological Explanation* (Nueva York, Random House, 1968), pág. 116.

mos eliminar la caracterización funcional en favor de una teoría que sea peculiar a alguno de tales sustratos. Esto excluiría nuestra posibilidad de describir la organización (abstracta) que una instanciación comparte con todas las demás. Una caracterización funcional de nuestros estados internos seguirá, por lo tanto, estando con nosotros.

Este segundo tema, como el primero, asigna un carácter ligeramente estipulativo a la PF, como si los sistemas empíricos tuvieran la responsabilidad de instanciar fielmente la organización que especifica la PF, en lugar de ser la PF la que tiene la responsabilidad de describir fielmente las actividades internas de una clase naturalmente distinta de sistemas empíricos. Esta impresión se ve realzada por los ejemplos estándar que se usan para ilustrar las pretensiones del funcionalismo —ratoneras, levanta válvulas, calculadoras aritméticas, computadoras, robots, etcétera. Éstos son artefactos construidos para satisfacer requisitos preconcebidos. En tales casos, un desajuste entre el sistema *físico* y la caracterización funcional pertinente sólo contraviene al primero y no a la última. La caracterización funcional resulta así sustraída de la crítica empírica, de un modo muy diferente al caso de una teoría empírica. Un funcionalista prominente —Hilary Putnam— ha argumentado con franqueza que la PF no es en modo alguno una teoría corregible.[8] Claramente, si la PF se construye sobre estos modelos, como suele ocurrir, es difícil que se plantee el problema de su integridad empírica, y menos aún, que reciba una respuesta crítica.

Si bien lo que antecede se ajusta a algunos funcionalistas, no se ajusta completamente a Fodor. Según su punto de vista, el objetivo de la psicología es encontrar la *mejor* caracterización funcional de nosotros mismos, y lo que ella sea constituye una cuestión empírica. También, su argumento a favor de la no eliminación del vocabulario mentalista de la psicología, no señala como ineliminable a la PF corriente, en particular. Sólo precisa sostener que se retenga *alguna* caracterización funcional abstracta, quizás alguna articulación o refinamiento de la PF.

Sin embargo, su apreciación del materialismo eliminativo sigue siendo baja. En primer lugar, es claro que Fodor piensa que en la PF no hay nada erróneo, en un sentido interesante o fundamental. Por el contrario, la concepción central de la actividad cognitiva de la PF —la manipulación de actitudes proposicionales— resulta ser el elemento central de la propia teoría de Fodor sobre la naturaleza del pensamien-

8. "Robots: Machines or Artificially Created Life?", *The Journal of Philosophy*, LXI, 21 (Nov. 12, 1964): 668-691, pp. 675, 681 y sigs.

to (véase *The Language of Thought*). Y en segundo lugar, queda pendiente la cuestión de que cualquiera que sea el arreglo que la PF pueda o no requerir, no puede ser desplazada por ninguna teoría naturalista de nuestro sustrato físico, puesto que son los rasgos funcionales abstractos de sus estados internos los que hacen a una persona, y no la química de su sustrato.

Todo esto resulta atractivo. Pero pienso que casi nada de ello es correcto. El funcionalismo ha disfrutado durante demasiado tiempo de una reputación osada y de *avant garde*. Es preciso que se le revele que es una posición miope y reaccionaria.

4. La naturaleza conservadora del funcionalismo

A partir del siguiente relato se puede obtener una perspectiva valiosa del funcionalismo. Para comenzar, recordemos la teoría de los alquimistas sobre la materia inanimada. Se trata, por supuesto, de una larga y abigarrada tradición y no de una única teoría. Pero una glosa será suficiente para nuestro propósito.

Los alquimistas concibieron lo "inanimado" como enteramente continuo con la materia animada, en el sentido de que las propiedades sensibles y conductuales de las diversas sustancias se deben a la animación [*ensoulment*] de la materia más baja por diversos espíritus o esencias. Estos aspectos no materiales, se sostenía, eran susceptibles de desarrollarse, tal como encontramos crecimiento y desarrollo en las almas diversas de las plantas, de los animales y de los humanos. La habilidad peculiar del alquimista consistía en saber cómo sembrar, nutrir y hacer madurar los espíritus deseables materializados en las combinaciones apropiadas.

Según una ortodoxia, los cuatro espíritus fundamentales (para la materia "inanimada") se denominaban "mercurio", "sulfuro", "arsénico amarillo" y "sal amoníaca". Se sostenía que cada uno de estos espíritus era responsable de un síndrome basto pero característico de las propiedades sensibles, combinatorias y causales. Por ejemplo, se sostenía que el espíritu mercurio era responsable de ciertos rasgos típicos de las sustancias metálicas: su brillantez, licuefactibilidad y otras. Se sostenía que el sulfuro era responsable de ciertos rasgos residuales típicos de los metales y de los rasgos exhibidos por la mena de la que podía destilarse el metal común. Cualquier sustancia metálica dada era, principalmente, una orquestación crítica de esos dos espíritus. Un relato similar valía

para los otros dos espíritus, y los cuatro tornaban inteligible y controlable un cierto dominio de rasgos y transformaciones físicos.

Por supuesto, el grado de control fue siempre limitado. O mejor aún, la predicción y control que poseían los alquimistas se debía más a la sabiduría manipulativa adquirida como aprendiz de un maestro, que a una comprensión genuina proporcionada por la teoría. La teoría seguía a la práctica, en lugar de dictarla. Pero la teoría proporcionaba cierto incentivo a la práctica y ante la ausencia de una alternativa desarrollada, era lo suficientemente convincente como para sostener una larga y tenaz tradición.

La tradición se había vuelto descolorida y fragmentada en la época en la que surge la química elemental de Lavoisier y Dalton para reemplazarla definitivamente. Pero supongamos que hubiera durado un poco más, quizá porque la ortodoxia de los cuatro espíritus se hubiera convertido en una parte remanida del sentido común, y examinemos la naturaleza del conflicto entre las dos teorías y algunas posibles vías de resolución.

Sin duda, la vía más simple de solución —la que históricamente tuvo lugar— es el reemplazo liso y llano.

La interpretación dualista de las cuatro esencias, como espíritus inmateriales, parecería irreflexiva e innecesaria, dado el poder de la taxonomía corpuscular de la química atómica. Y una reducción de la vieja taxonomía a la nueva parecería imposible, en la medida en que la vieja teoría, comparativamente impotente, clasificara a las cosas de manera distinta [*cross-classifies things*] a como lo hace la nueva teoría. La eliminación aparecería entonces como la única alternativa, a *menos que* algún astuto y decidido defensor de la visión alquímica tuviera talento como para sugerir la siguiente defensa.

Ser "animado por mercurio" o por "sulfuro" o por cualquiera de los otros dos espíritus, es en realidad un estado *funcional*. Por ejemplo, el primero se define por la disposición a reflejar la luz, a licuarse con el calor, a unirse con otra materia en el mismo estado, etcétera. Y cada uno de estos cuatro estados está relacionado con los otros de modo tal que el síndrome de cada uno de ellos varía en función del otro estado, también instanciado en el mismo sustrato. Así el nivel de descripción abarcado por el vocabulario alquímico es abstracto: distintas sustancias materiales, adecuadamente "animadas", pueden exhibir los rasgos de un metal o aún específicamente los del oro. Porque es el síndrome total de las propiedades causales efectivamente dadas [*occurrent*] lo que importa, y no los detalles corpusculares del sustrato. La alquimia, se concluye,

abarca un nivel de organización de la realidad distinto de la organización que se da en el nivel de la química corpuscular, e irreductible a ella.

Este punto de vista podría haber tenido una atracción considerable. Después de todo, evita a los alquimistas el peso de tener que defender ánimas inmateriales que van y vienen; los libera de tener que enfrentarse a las durísimas exigencias de una reducción naturalista; les evita el conflicto y la confusión de la eliminación lisa y llana. ¡La teoría alquimista aparece como básicamente correcta! Los alquimistas no necesitan aparecer tampoco como demasiado obstinados o dogmáticos en esto. La alquimia, tal como está —conceden— puede requerir una reorganización sustancial, y la experiencia tiene que ser nuestra guía. Pero, nos recuerdan, no tenemos que temer su reemplazo naturalista puesto que es la orquestación particular de los síndromes de propiedades causales efectivamente dadas la que hace oro a un trozo de materia, no los detalles idiosincrásicos de su sustrato corpuscular. Una circunstancia ulterior habría vuelto aún más plausible a esta alegación. Porque el hecho es que los alquimistas *realmente* sabían cómo hacer oro, en este sentido de 'oro', pertinentemente debilitado, y podían hacerlo de variadas maneras. Su "oro" nunca era, ¡ay!, tan perfecto como el "oro" nutrido en la matriz de la naturaleza, pero, ¿qué mortal puede esperar competir con la naturaleza misma?

Lo que este relato muestra es que es posible, al menos, que la constelación de movidas, pretensiones y defensas características del funcionalismo constituya un atropello a la razón y a la verdad, y que lo sea con una plausibilidad aterradora. La alquimia es una teoría mala, que bien merece su completa eliminación, y la defensa que acabamos de explorar es reaccionaria, oscurantista, retrógrada y errónea. Pero, en el contexto histórico, aun para la gente razonable, esa defensa podría haber parecido completamente sensata.

El ejemplo de la alquimia es un caso deliberadamente claro de lo que bien podría denominarse "la estratagema funcionalista", y no es difícil imaginar otros casos. Siguiendo estos lineamientos también se puede construir una defensa extrema de la teoría del flogisto: interprétese como estados funcionales el estar muy flogistizado y el estar deflogistizado, definiéndolos en términos de ciertos síndromes de disposiciones causales; señálese la gran variedad de sustratos naturales capaces de combustión y calcificación; aléguese una integridad funcional irreductible para lo que ha probado carecer de integridad natural, y déjense a un lado en los defectos restantes, bajo la promesa de idear mejoras futuras. Una receta similar proporcionará nueva vida a los cuatro humores

de la medicina medieval, a la esencia vital o al principio de la biología pre-moderna, etcétera.

Si la aplicación de la estratagema funcionalista a estos otros casos sirve de algo, [es claro que] resulta ser una cortina de humo para preservar el error y la confusión. ¿De dónde emana nuestra certidumbre de que en las revistas especializadas contemporáneas no se está jugando una charada similar en nombre de la PF? El paralelismo con el caso de la alquimia es en todos sus aspectos penosamente completo, ¡tal como lo es el paralelismo entre la búsqueda del oro artificial y la búsqueda de la inteligencia artificial!

No se me malentienda en relación con este último punto. Ambos son objetivos respetables: gracias a la física nuclear el oro artificial (pero real) está finalmente a nuestro alcance, aunque sólo en cantidades submicroscópicas, y la inteligencia artificial (pero real) eventualmente lo estará. Pero, así como la instrumentación cuidadosa de síndromes superficiales para producir oro genuino, fue el modo equivocado de producirlo, la instrumentación cuidadosa de síndromes superficiales puede ser el modo equivocado de producir inteligencia genuina. Como con el oro, lo que se puede requerir es que la ciencia penetre en la clase *natural* subyacente que da origen, de modo directo, al síndrome total.

En síntesis, cuando enfrentamos a la impotencia explicativa, la historia estancada y el aislamiento sistemático de las locuciones [*idioms*] intencionales de la PF, insistir en que tales locuciones son abstractas, funcionales y de carácter irreductible, no constituye una respuesta adecuada o interesante. En primer lugar, esta misma defensa podría haber sido armada con aceptable plausibilidad sin que importe *qué* red desordenada de estados internos nos ha adscripto nuestro folklore. Y en segundo lugar, la defensa supone esencialmente lo que está en cuestión: supone que son las locuciones intencionales de la PF, poco más o menos, las que expresan los rasgos *importantes* que comparten todos las sistemas cognitivos. Pero pueden no hacerlo. Por cierto que es erróneo suponer que lo hacen y argumentar luego en contra de la posibilidad de un reemplazo materialista, sobre la base de que tiene que describir cuestiones en un nivel que es diferente del nivel importante. Esto es, precisamente, una petición de principio en favor del marco más antiguo.

Finalmente, es importante señalar que el materialismo eliminativo es *consistente* con la alegación de que la esencia de un sistema cognitivo reside en la organización funcional abstracta de sus estados internos. El materialista eliminativo no está comprometido con la idea de que la descripción correcta de la cognición *tenga que* ser una descripción natura-

lista, si bien puede perdonársele explorar esa posibilidad. Lo que sostiene es que la descripción correcta de la cognición, ya sea funcionalista o naturalista, guardará tanta semejanza con la PF como la química moderna guarda con la alquimia de los cuatro espíritus.

Tratemos de encarar ahora el argumento contra el materialismo eliminativo, que se centra en la dimensión normativa de la PF. Creo que podemos hacerlo de manera bastante rápida.

Primero, el hecho de que las regularidades adscriptas por el núcleo intencional de la PF sean predicadas de ciertas relaciones lógicas entre proposiciones, no da en sí mismo un fundamento para alegar algo esencialmente normativo respecto de la PF. Para trazar un paralelo pertinente, el hecho de que las regularidades adscriptas por la ley clásica de los gases sean predicadas de relaciones aritméticas entre números, no implica nada esencialmente normativo acerca de la ley clásica de los gases. Y las relaciones lógicas entre proposiciones son tanto una cuestión objetiva acerca de un hecho abstracto, como lo son las relaciones aritméticas entre números. En este respecto, la ley

(4) (x) (p) (q) $(((x$ cree que $p)$ & $(x$ cree que (si p entonces $q))) \supset$ (salvo confusión, distracción etc., x cree que q))

está a la par de la ley clásica de los gases,

(6) (x) (P) (V) (μ) $(((x$ tiene una presión $P)$ & $(x$ tiene un volumen $V)$ & $(x$ tiene una cantidad $\mu)) \supset$ (salvo una presión o densidad muy alta, x tiene una temperatura de PV / μ R).

La dimensión normativa entra sólo porque *valoramos* la mayoría de las pautas adscriptas por la PF. Pero no las valoramos a todas. Considérese,

(7) (x) (P) $(((x$ desea con todo su corazón que $p)$ & $(x$ llega a saber que $—p)) \supset$ (salvo una fuerza inusual de carácter, x se conduele de que $—p$)).

Más aún, y tal como generalmente ocurre con la convicciones normativas, una intuición nueva puede provocar grandes cambios en lo que valoramos.

Segundo, las leyes de la PF nos adscriben sólo una racionalidad mínima y truncada, no una racionalidad ideal, como algunos han sugerido. La racionalidad caracterizada por el conjunto de todas las leyes de

la PF no llega a ser una racionalidad ideal. Esto no es sorprendente. Por lo demás, no tenemos una concepción clara o acabada de la racionalidad ideal. Por cierto que el hombre común tampoco la tiene. Por tal razón, no es plausible suponer que los fracasos explicativos que padece la PF se deban primordialmente al fracaso humano de vivir de acuerdo con el patrón ideal que ella proporciona. Muy por el contrario, la concepción de la racionalidad que proporciona parece floja y superficial, especialmente cuando se la compara con la complejidad dialéctica de nuestra historia científica o con el virtuosismo raciocinativo que exhibe cualquier niño.

Tercero, aun si nuestra concepción corriente de la racionalidad y, más generalmente, de la virtud cognitiva, está constituida en gran medida dentro del marco oracional/proposicional de la PF, no hay garantía de que ese marco sea adecuado para la descripción más profunda y más precisa que se necesita. Aun si concedemos la integridad categorial de la PF, al menos cuando se la aplica a los humanos en tanto usuarios del lenguaje, no es nada claro que los parámetros de la virtud intelectual tengan que hallarse en el nivel categorial abarcado por las actitudes proposicionales. Después de todo, el uso del lenguaje es algo que es aprendido por un cerebro que ya es capaz de una vigorosa actividad cognitiva. El uso del lenguaje es adquirido como una destreza más entre una gran variedad de destrezas manipulativas aprendidas, y es conducido por un cerebro moldeado por la evolución para realizar una gran cantidad de funciones, siendo el uso del lenguaje sólo la más reciente y, quizá, la última de ellas. Contra el trasfondo de estos hechos, el uso del lenguaje aparece como una actividad extremadamente periférica, como un modo de interacción social específico de la especie [*species specific*] que es dominado gracias a la versatilidad y al poder de un modo de actividad más básico. ¿Por qué aceptar entonces una teoría de la actividad cognitiva que modela sus elementos sobre los elementos del lenguaje humano? ¿Y por qué suponer que los parámetros fundamentales de la virtud intelectual son o pueden ser definidos a partir de los elementos de ese nivel superficial?

De tal modo, un avance serio en nuestra apreciación de la virtud cognitiva parecería *requerir* que vayamos más allá de la PF, que superemos la pobreza de la concepción de la racionalidad de la PF, trascendiendo completamente su cinemática proposicional, desarrollando una cinemática más profunda y más general de la actividad cognitiva y distinguiendo, en ese nuevo marco, qué modos de actividad cinemáticamente posibles tienen que ser valorados y estimulados (como más efi-

cientes, confiables, productivos, etcétera). El materialismo eliminativo no implica, así, el fin de nuestras preocupaciones normativas. Sólo implica que ellas tendrán que ser reconstituidas en un nivel de comprensión más revelador: el nivel que va a proporcionar una neurociencia madura [*mature*].

Exploraremos ahora lo que un futuro teoréticamente ilustrado podría reservarnos. No porque podamos preverlo con alguna claridad especial, sino porque es importante romper el dominio que la cinemática proposicional de la PF ejerce sobre nuestra imaginación. En lo que respecta a la presente sección, podemos resumir nuestras conclusiones como sigue. La PF no es nada más ni nada menos que una teoría culturalmente protegida acerca de cómo nosotros y los animales superiores funcionamos. No tiene rasgos esenciales que la hagan empíricamente invulnerable, ni funciones únicas que la hagan irreemplazable, ni condiciones especiales de ninguna clase. Prestemos, pues, oídos escépticos a cualquier alegato especial en su nombre.

5. *Más allá de la psicología* folk

¿Qué podría involucrar, realmente, la eliminación de la PF: no sólo las locuciones, comparativamente directas, que corresponden a las sensaciones, sino el aparato completo de las actitudes proposicionales? Ello depende, principalmente, de lo que pueda descubrir la neurociencia y de nuestra determinación a capitalizarlo. He aquí tres escenarios en los que la concepción operativa de la actividad cognitiva es progresivamente divorciada de las formas y de las categorías que caracterizan al lenguaje natural. Si el lector consiente la falta de sustancia real, intentaré delinear una forma plausible.

Primero, supongamos que la investigación de la estructura y actividad del cerebro, tanto de grano fino como global, produce una nueva cinemática y una dinámica correlativa para lo que ahora se considera que es la actividad cognitiva. La teoría es uniforme para todos los cerebros terrestres, no sólo los cerebros humanos, y hace contactos conceptuales apropiados con la biología evolutiva y la termodinámica del no equilibrio [*non-equilibrium thermodynamics*]. Nos adscribe, en todo momento, un conjunto o configuración de estados complejos que están especificados dentro de la teoría como "sólidos" figurativos dentro de un espacio de fases tetra o pentadimensionales. Las leyes de la teoría gobiernan la interacción, el movimiento y la transformación de esos

estados "sólidos" dentro de ese espacio y también sus relaciones con cualesquiera de los transductores sensorios y motores que el sistema posea. Como con la mecánica celeste, no es posible en la práctica, por muchas razones, especificar con exactitud a los "sólidos" involucrados y describir exhaustivamente todos los "sólidos" adyacentes dinámicamente relevantes, pero también resulta aquí que las aproximaciones obvias a las que recurrimos producen excelentes explicaciones/predicciones del cambio interno y de la conducta externa, por lo menos en el corto plazo. Respecto de la actividad a largo plazo, la teoría proporciona explicaciones poderosas y unificadas del proceso de aprendizaje, de la naturaleza de la enfermedad mental y de las variaciones en el carácter y en la inteligencia, tanto para el reino animal como para los individuos humanos.

Más aún, la teoría proporciona una explicación directa del "conocimiento", tal como es concebido tradicionalmente. De acuerdo con la nueva teoría toda oración declarativa a la cual un hablante prestara asentimiento sería meramente una *proyección* unidimensional —a través de la lente compuesta por las áreas de Wernicke y de Broca en la superficie idiosincrásica del lenguaje del hablante [*the idiosyncratic surface of the speaker's language*]—, una proyección unidimensional de un "sólido" tetra o penta-dimensional que es un elemento en el verdadero estado cinemático de tal hablante. (Recuérdense las sombras en la pared de la caverna de Platón.) Al ser proyecciones de esa realidad interna, tales oraciones portan información significativa respecto de ella y de tal modo son adecuadas para funcionar como elementos en un sistema de comunicación. Por otra parte, al ser proyecciones *sub*dimensionales, sólo reflejan una parte restringida de la realidad proyectada. Por lo tanto, *no* son adecuadas para representar la realidad más profunda, en todos sus aspectos cinemática, dinámica y aun normativamente pertinentes. Es decir, es inexorable que un sistema de actitudes proposicionales tal como la PF, no capture lo que está ocurriendo allí, aunque pueda reflejar suficiente estructura superficial como para convalidar una tradición de tipo alquimista entre quienes carecen de una teoría mejor. Sin embargo, desde la perspectiva de la teoría más nueva, es claro que no existen estados gobernados por leyes de la clase que postula la PF. Las leyes reales que gobiernan nuestras actividades internas se definen sobre estados y configuraciones cinemáticas diferentes y mucho más complejas, tal como son [definidos] los criterios normativos para la integridad del desarrollo y la virtud intelectual.

Un resultado teórico del tipo que se acaba de describir puede ser

considerado, adecuadamente, como un caso de eliminación de una ontología teórica en favor de otra, pero el éxito aquí imaginado por la neurociencia sistemática [*systematic*] no necesita tener ningún efecto apreciable sobre la práctica común. Los viejos estilos difícilmente mueren y en ausencia de alguna necesidad práctica, pueden no morir nunca. Aún así, no es inconcebible que algún segmento de la población, o toda ella, llegue a familiarizarse íntimamente con el vocabulario requerido para caracterizar nuestros estados cinemáticos, aprender las leyes que gobiernan sus interacciones y proyecciones conductuales, adquirir cierta habilidad para adscripciones en primera persona y desplazar completamente el uso de la PF, aun en la plaza pública. Entonces, el deceso de la ontología de la PF sería total.

Ahora podemos explorar una segunda posibilidad, bastante más radical. Todos estamos familiarizados con la tesis de Chomsky de que la mente o el cerebro humano contiene de modo innato y único las estructuras abstractas para aprender y usar los lenguajes naturales específicamente humanos. Una hipótesis alternativa es que nuestro cerebro contiene efectivamente estructuras innatas pero que esas estructuras tienen como función original y aun primordial la organización de la experiencia perceptual, siendo la administración de categorías lingüísticas una función adquirida y adicional para la cual la evolución las ha adaptado sólo incidentalmente.[9] Esta hipótesis tiene la ventaja de no requerir el salto evolutivo que el enfoque de Chomsky parecería requerir, y tiene además otras ventajas. Pero esas cuestiones no necesitan preocuparnos aquí. Supongamos, para nuestros propósitos, que este punto de vista alternativo es verdadero y consideremos el siguiente relato.

La investigación de las estructuras neurales que dan base a la organización y al procesamiento de la información perceptual, revela que son capaces de conducir una gran variedad de tareas complejas, algunas de las cuales muestran una complejidad muchísimo mayor que la que exhibe el lenguaje natural. Resulta ser que los lenguajes naturales sólo explotan una porción muy elemental de la maquinaria disponible, el grueso de la cual sirve para actividades mucho más complejas, más allá del alcance de las concepciones proposicionales de la PF. La aclaración detallada de lo que es esa maquinaria y de las capacidades que tiene torna evidente que una forma de lenguaje mucho más sofisticada que

9. Richard Gregory defiende una opinión similar en "The Grammar of Vision", *Listener*, XXXIII, 2133 (febrero de 1970): 242-246; reimpreso en su *Concepts and Mechanisms of Perception* (Londres, Duckworth, 1975), 622-629.

el lenguaje "natural", aunque decididamente "extraña" a sus estructuras sintácticas y semánticas, también podría ser aprendido y usado por nuestros sistemas innatos. Se comprende de inmediato que tal sistema novedoso de comunicación podría elevar enormemente la eficiencia del intercambio de información entre cerebros, y aumentaría la evaluación cognitiva en una magnitud comparable, puesto que reflejaría la estructura subyacente de nuestras actividades cognitivas con mayor detalle que como lo hace el lenguaje natural.

Guiados por esta nueva comprensión de esas estructuras internas, logramos construir un nuevo sistema de comunicación verbal, totalmente diferente del lenguaje natural, con una gramática combinatoria nueva y más poderosa aplicada a elementos novedosos que forman combinaciones novedosas, con propiedades exóticas. Las secuencias [*strings*] compuestas de este sistema alternativo, llamadas "*Ubersätzen*", no son evaluadas como verdaderas o falsas, ni las relaciones entre ellas son en modo alguno análogas a las relaciones de implicación, etcétera, que valen entre las oraciones. Ellas muestran una organización diferente y manifiestan virtudes diferentes.

Una vez construido, este "lenguaje" prueba ser aprendible, tiene el poder que se ha proyectado y en dos generaciones ha barrido el planeta. Todos usan el nuevo sistema. Las formas sintácticas y las categorías semánticas del así llamado lenguaje "natural", desaparecen totalmente. Y con ellas desaparecen las actitudes proposicionales de la PF, desplazadas por un esquema más revelador en el cual, por supuesto, las actitudes "ubersätzenales" ["*ubersätzenal attitudes*"] desempeñan el papel principal. Nuevamente la PF quedaría eliminada.

Nótese que este segundo relato ilustra un tema con infinitas variaciones. Hay tantas posibles "psicologías de sentido común" diferentes, como hay posibles sistemas de comunicación diferentemente estructurados que les sirven de modelo.

Una tercera y aún más extraña posibilidad puede describirse como sigue. Sabemos que existe una considerable lateralización de función entre los dos hemisferios cerebrales y que ambos hemisferios hacen uso de la información que obtienen entre sí mediante la gran comisura cerebral, el cuerpo calloso, un cable gigante de neuronas que los conecta. Los pacientes cuya comisura ha sido seccionada quirúrgicamente exhiben una variedad de carencias conductuales que indican una pérdida de acceso a la información que habitualmente un hemisferio obtenía del otro. Sin embargo, en las personas con agénesis callosa [*callosal agenesis*] (un defecto congénito en el que el cable conector simplemente

está ausente), hay poca o ninguna deficiencia conductual, lo cual sugiere que los dos hemisferios han aprendido a explotar la información transportada por vías menos directas, vías que los conectan a través de las regiones subcorticales. Esto sugiere que aun en el caso normal, un hemisferio en desarrollo *aprende* a hacer uso de la información que la comisura cerebral deposita en su puerta de entrada. Lo que tenemos, entonces, en el caso del ser humano normal, son dos sistemas cognitivos físicamente distintos (ambos capaces de funcionar independientemente) que responden de un modo sistemático y aprendido a la información intercambiada. Y lo que es especialmente interesante respecto de este caso es el monto de la información intercambiada. El cable de la comisura consiste en aproximadamente 200 millones de neuronas,[10] y aun si supusiéramos que cada una de esas fibras es capaz de uno de sólo dos estados posibles por segundo (una estimación muy conservadora), estamos considerando un canal cuya capacidad de información es mayor que 2×10^8 bits binarios por segundo. Compárese esto con la capacidad de menos de 500 bits por segundo del inglés oral.

Ahora bien, si dos hemisferios distintos pueden aprender a comunicarse en una escala tan impresionante, ¿por qué no podrían aprenderlo también dos cerebros distintos? Esto requeriría una "comisura" artificial de algún tipo, pero permítasenos suponer que podemos diseñar un transductor factible para su implantación en el lugar del cerebro que la investigación revele que es conveniente, un transductor que convierta una sinfonía de actividad neural en (digamos) microondas emitidas desde una antena en la frente, y que realice la función inversa de convertir en actividad neural las microondas recibidas. Conectarlo no tiene por qué ser necesariamente un problema insuperable. Simplemente, burlamos a los procesos normales de la arborización dendrítica para que hagan crecer su propia miríada de conexiones en la microsuperficie activa del transductor.

Una vez que se ha abierto el canal entre dos o más personas, ellas pueden aprender (*aprender*) a intercambiar información y a coordinar sus conductas con la misma intimidad y virtuosismo que los que exhiben nuestros propios hemisferios cerebrales. ¡Piénsese lo que esto podría significar para equipos de hockey, compañías de ballet y equipos de investigación! Si la población completa estuviera así equipada, el lenguaje oral de cualquier tipo podría muy bien desaparecer completamen-

10. M. S. Gazzaniga y J. E. LeDoux, *The Integrated Mind* (Nueva York, Plenum Press, 1975).

te, víctima del principio "por qué arrastrarte si puedes volar". Las bibliotecas se llenarían, no con libros sino con largos registros de intercambios ejemplares de actividad neural. Ellas constituirían una herencia cultural creciente, un "Tercer Mundo" en evolución, para usar la terminología de Karl Popper. Pero no consistirían de oraciones o de argumentos.

¿Cómo comprenderán y concebirán tales personas a los demás individuos? A esta pregunta sólo puedo contestar: "Aproximadamente de la misma manera en que su hemisferio derecho 'comprende' y 'concibe' su hemisferio izquierdo, ¡íntima y eficientemente pero no proposicionalmente!".

Estas especulaciones, espero, evocarán un sentido adecuado de posibilidades inexploradas; en todo caso, las daré por terminadas aquí. Su función es producir algunas incursiones en el aura de inconcebibilidad que habitualmente circunda la idea de que podríamos rechazar la PF. La tensión conceptual experimentada encuentra expresión incluso en un argumento a favor de la tesis de que el materialismo eliminativo es incoherente porque niega las condiciones mismas asumidas en el supuesto de que es significativo. Terminaré con una breve discusión de esta popular movida.

Tal como se la formula, la *reductio* procede señalando que la enunciación del materialismo eliminativo no es más que una cadena no significativa de marcas o ruidos, a menos que tal cadena sea la expresión de una cierta *creencia*, de una cierta *intención* de comunicar, de un *conocimiento* de la gramática del lenguaje, etcétera. Pero si el enunciado del materialismo eliminativo es verdadero, entonces no hay estados tales que expresar. El enunciado en cuestión sería entonces una cadena no significativa de marcas o ruidos. Por lo tanto *no* sería verdadero. En consecuencia, no es verdadero. Q.E.D.

La dificultad con cualquier *reductio* no formal es que la conclusión en contra de la suposición inicial nunca es mejor que las suposiciones materiales invocadas para alcanzar la conclusión incoherente. En este caso, las suposiciones adicionales involucran una cierta teoría del significado, una teoría que presupone la integridad de la PF. Pero hablando formalmente, uno podría también inferir, a partir del resultado incoherente, que es esa teoría del significado la que debe ser rechazada. Dada la crítica independiente formulada antes a la PF, ésta pareciera la opción preferible. Pero, de cualquier manera, uno no puede suponer simplemente esa teoría particular del significado sin hacer una petición de principio con respecto a la integridad de la PF.

La naturaleza circular de este argumento está gráficamente ilustrada por la siguiente analogía que debo a Patricia Churchland.[11] La cuestión ahora, ubicada en el siglo XVII, es si existe o no una sustancia tal como el *espíritu vital*. En esa época se sostenía que esa sustancia distinguía lo animado de lo inanimado, sin que existiera un reconocimiento significativo de alternativas reales. Dado el monopolio ejercido por esa concepción, dado el grado con que estaba integrada con muchas otras de nuestras concepciones y dada la magnitud de las revisiones que cualquiera otra concepción alternativa sería requeriría, la siguiente refutación de cualquier pretensión antivitalista, se consideraría instantáneamente plausible.

> El antivitalista dice que no hay tal cosa como el espíritu vital. Pero esta pretensión se autorrefuta. El hablante puede esperar que se lo tome en serio sólo si su pretensión no lo es. Porque si la alegación es verdadera, entonces el hablante no tiene espíritu vital y debe estar *muerto*. Pero si está muerto, entonces su enunciado es una cadena no significativa de ruidos, vacía de razón y de verdad.

La naturaleza circular de este argumento no exige elaboración, supongo. Recomiendo a quienes estén impresionados por el argumento anterior, que examinen el paralelo.

La tesis de este artículo puede ser resumida como sigue. Las actitudes proposicionales de la psicología *folk* no constituyen una barrera infranqueable para la marea progresiva de la neurociencia. Por el contrario, el desplazamiento razonado de la psicología *folk* no es sólo altamente posible, sino que representa uno de los desplazamientos teóricos más estimulantes que podamos imaginar en la actualidad.

TRADUCTORAS: Ana C. Couló, María C. González y Nora Stigol.
REVISIÓN TÉCNICA: Eduardo Rabossi.

11. "Is Determinism Self-Refuting?", *Mind*, 90 (1981), págs. 99-101.

CAPÍTULO 3

LA PERSISTENCIA DE LAS ACTITUDES *

Jerry A. Fodor

Sueño de una noche de verano, acto 3, escena 2 **

Entran Demetrio y Hermia.
Dem.: ¡Oh! ¿Por qué rechazáis a quien os ama con tanto ardor? Regañad [a] quien os deteste, mas no [a] quien os adora.
Herm.: No te hago sentir más que mis desdenes, cuando podría tratarte peor, porque temo que me has dado motivos para maldecirte. Si es verdad que has muerto a Lisandro mientras se hallaba dormido, acaba, ya que tienes un pie en el crimen, acaba de hundirte en él y mátame igualmente. No es el sol más fiel al día que Lisandro a mí. ¿Puedo creer que haya abandonado a Hermia dormida? Antes creería que la Tierra puede atravesarse de parte a parte, y que la Luna, penetrando a través de su centro hasta los antípodas, podría venir en pleno mediodía a perturbar los rayos de su hermano. Imposible es que no le hayas dado muerte. Tu cara, feroz y siniestra, es, sin duda, la de un asesino.

Muy bonito. Y también muy *plausible*. Un ejemplo convincente (aunque informal) de inferencia teórica implícita, no demostrativa.

He aquí cómo debe de haber sido la inferencia, dejando a un lado un montón de proposiciones subsidiarias [*lemmas*]: Hermia tiene razones para creer que Lisandro la ama. (Lisandro le ha dicho que la ama —varias veces y en elegantes versos yámbicos— y las inferencias de lo que la gente efectivamente siente, a partir de lo que dice que siente, son confiables, *ceteris paribus*.) Pero si Lisandro realmente ama a Hermia, entonces, a fortiori, le desea lo mejor. Y si Lisandro desea a Hermia lo mejor, entonces Lisandro no la abandona voluntariamente en un bosque oscuro. (Puede haber leones. "No hay fiera más terrible que un león

* Capítulo 1 de *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA, MIT Press, 1987, págs. 1-26. Con autorización del autor y de MIT Press.
** Traducción de Luis Astrana Marín. Shakespeare: *Obras Completas*, Aguilar, Madrid, 1974.

vivo".) Pero, de hecho, Hermia fue abandonada de esa manera por Lisandro. Luego, no fue adrede. Luego, es plausible pensar que Lisandro haya sufrido algún daño. ¿A manos de quién? Presumiblemente, a manos de Demetrio. Porque Demetrio es el rival de Lisandro en el amor de Hermia, y se presume que los rivales en amor *no* se quieren bien. Específicamente, Hermia cree que Demetrio cree que Lisandro vivo es un obstáculo para el éxito de sus insinuaciones (las de Demetrio a Hermia). Más aún, Hermia cree (correctamente) que si x quiere que P, y x cree que no P a menos que Q, y x cree que x puede producir Q, entonces (*ceteris paribus*) x trata de producir Q. Más aún, Hermia cree (de nuevo, correctamente) que por lo general la gente logra producir lo que trata de producir. *Luego*: sabiendo y creyendo todo esto, Hermia infiere que quizá Demetrio ha matado ya a Lisandro. Y nosotros, la audiencia, que sabemos lo que Hermia sabe y cree, y que compartimos más o menos sus puntos de vista sobre la psicología de amantes y de rivales, comprendemos cómo ha llegado a hacer esa inferencia. Nos condolemos.

De hecho, Hermia está completamente equivocada. Demetrio es inocente y Lisandro vive. La intrincada teoría que conecta creencias, deseos y acciones —la teoría implícita en la que se apoya Hermia para entender qué hizo Lisandro y qué puede haber hecho Demetrio, y en la que *nosotros* confiamos para entender por qué Hermia infiere lo que infiere, y en la que Shakespeare se apoya para predecir y manipular nuestras simpatías (entre paréntesis, "*desconstrucción*", *un bledo*)— no tiene en cuenta las intervenciones nocturnas de hadas traviesas. Sin que Hermia lo sepa, un duende peripatético ha generado la cláusula *ceteris paribus* y ha hecho que su plausible inferencia fracase. "La razón y el amor se visitan poco ahora: es peor para todos que algunos honestos vecinos no logren amigarlos."

Aceptando, sin embargo, que la teoría falla de vez en cuando —y no sólo cuando intervienen las hadas— de todas maneras quiero enfatizar (1) *cuán a menudo funciona bien*; (2) *qué profunda es*; y (3) *en qué gran medida dependemos de ella*. La psicología de creencias/deseos de sentido común [*commonsense belief/desire psychology*] ha sido sometida, recientemente, a una gran presión filosófica, y es posible poner en duda que se la pueda salvar, dados los tipos de problemas que sus críticos han planteado. Hay, sin embargo, una cuestión previa: la de si vale la pena hacer el esfuerzo de salvarla. Éste es el tema con el que me propongo comenzar.

1. Cuán a menudo funciona bien

Hermia se equivocó: su amante era menos constante de lo que había supuesto. Las aplicaciones de la psicología de sentido común [*commonsense psychology*] median nuestras relaciones con los demás, y cuando sus predicciones fallan, esas relaciones se rompen. Es muy probable que la confusión resultante suceda en público y llame mucho la atención.

> *Herm.*: En el corto espacio de una noche me habéis amado y me habéis dejado. ¡Me habéis dejado! ¡Oh! ¡Los dioses me libren de creerlo! ¿Es de veras?
> *Lis.*: Sí, ¡por mi vida!, y con la firme intención de no volver a verte. Desecha en cuanto a eso toda especie de esperanza...

Esta clase de cosas da lugar a un excelente teatro; los *éxitos* de la psicología de sentido común, por contraposición, son ubicuos y —por esa misma razón— prácticamente invisibles.

La psicología de sentido común funciona tan bien que no se nota. Es como esos míticos Rolls Royce cuyos motores se sellan en la fábrica; sólo que es mejor porque no es mítica. Alguien que no conozco llama por teléfono a mi oficina en Nueva York desde —por ejemplo— Arizona. '¿Querría dar una conferencia aquí el próximo martes?', son las palabras que pronuncia. 'Sí, gracias. Llegaré al aeropuerto en el vuelo de las 15', son las palabras con las que respondo. Eso es *todo* lo que sucede, pero es más que suficiente; la carga de predecir la conducta —de salvar el hiato entre proferencias [*utterances*] y acciones— recae habitualmente en la teoría. Y la teoría funciona tan bien que varios días después (o semanas después, o meses después, o años después; uno puede variar el ejemplo a gusto) y varios miles de millas más lejos, ahí estoy en el aeropuerto, y allí está él esperándome. Y si yo *no* aparezco, es más probable que algo haya andado mal en la aerolínea y no que haya fallado la teoría. No es posible estimar, en términos cuantitativos, el éxito con que la psicología de sentido común nos permite coordinar nuestras conductas. Pero tengo la impresión de que nos las arreglamos bastante bien unos con otros; a menudo bastante mejor de lo que nos las arreglamos con máquinas menos complejas.

El punto —una vez más— es que la teoría de la que recibimos ese extraordinario poder predictivo es, precisamente, la vieja y querida psicología de creencias/deseos de sentido común. Ella es la que nos dice, por ejemplo, cómo inferir las intenciones de la gente a partir de los soni-

dos que produce (si alguien emite la forma lingüística [*form of words*] 'Llegaré al aeropuerto en el vuelo de las 15 hs.', entonces, *ceteris paribus*, tiene la intención de llegar al aeropuerto en el vuelo de las 15), y cómo inferir la conducta de la gente a partir de sus intenciones (si alguien tiene la intención de llegar al aeropuerto en el vuelo de las 15, entonces, *ceteris paribus*, se comportará de modo tal que llegará a ese lugar a esa hora, a menos que se den fallas mecánicas o actos divinos). Y todo esto no sólo funciona con personas cuya psicología uno conoce íntimamente: nuestros amigos más cercanos, digamos, o nuestra esposa del alma. Funciona con *extraños*, con gente que uno no reconocería aunque tropezara con ella. Y no sólo funciona en condiciones de laboratorio —donde uno puede controlar las variables en interacción— sino también, de manera prominente, en condiciones de campo en las que todo lo que uno sabe sobre el origen de las variaciones es lo que la psicología de sentido común nos dice sobre ellas. Es notable. Si tuviéramos el mismo éxito con las predicciones del clima, nadie se mojaría los pies; y sin embargo la etiología del clima tiene que ser seguramente un juego de niños comparada con las causas de la conducta.

Sí, ¿pero, qué pasa con todas esas cláusulas *ceteris paribus*? Haré una digresión.

Los filósofos argumentan a veces que la apariencia de adecuación predictiva que rodea a las generalizaciones de la psicología de sentido común, es espuria. Porque —dicen— tan pronto como uno trata de explicitar esas generalizaciones, ve que tienen que ser complementadas con cláusulas *ceteris paribus*; complementadas de tal modo que se las hace no refutables de una manera *trivial*. "Falsa o vacua" es la acusación.

Consideremos la revocabilidad [*defeasibility*] de 'Si alguien emite la forma lingüística "Llegaré al aeropuerto en el vuelo de las 15", entonces tiene la intención de estar en el aeropuerto a las 15' Esta generalización *no* vale si, por ejemplo, el hablante miente; o si el hablante está usando la proferencia como un ejemplo (de una oración falsa, digamos); o si es un hablante monolingüe del urdu que ha proferido esa oración accidentalmente, o si el hablante está hablando en sueños; o... lo que sea. Por supuesto, uno puede defender la generalización del modo habitual; puede decir que '*en igualdad de circunstancias* [*all else being equal*], si alguien emite la forma lingüística "Llegaré al aeropuerto en el vuelo de las 15", entonces tiene la intención de estar en el aeropuerto en el vuelo de las 15' Pero quizás esto último no significa más que: 'Si alguien dice que tiene la intención de estar allí, entonces realmente tiene la intención

de estar allí, a menos que no la tenga'. Por supuesto que con toda seguridad, eso es predictivamente adecuado; nada de lo que suceda lo refutará; nada de lo que sucediera podría refutarlo.

Muchos filósofos parecen estar impresionados por este tipo de argumento; sin embargo, aun a primera vista, sería sorprendente que sirviera de algo. Después de todo, usamos las generalizaciones psicológicas de sentido común para predecir nuestras respectivas conductas, y las predicciones —muy a menudo— resultan verdaderas. Pero, ¿cómo podrían serlo si las generalizaciones en las que basamos las predicciones fueran *vacuas*?

Me inclino a pensar que lo que se alega acerca de la dependencia implícita de la psicología de sentido común respecto de las cláusulas *ceteris paribus* no efectivizadas, es de hecho una propiedad usual de las generalizaciones *explícitas* en *todas* las ciencias especiales; vale decir, en todos los esquemas explicativos empíricos, excepto en la física básica. Considérese la siguiente modesta verdad de la geología: un río con meandros erosiona su margen. "Falso o vacuo", podría argüir un filósofo. "Tómesela al pie de la letra —como una generalización universal estricta— y seguramente es falsa. Piénsese en el caso en que el clima cambia y el río se hiela; o el mundo se acaba; o alguien construye un dique; o alguien construye una pared de concreto en la margen externa; o las lluvias cesan y el río se seca... o lo que sea. Uno puede, por supuesto, defender la generalización de la manera habitual, añadiendo una cláusula *ceteris paribus*: '*En igualdad de circunstancias*, un río con meandros erosiona su margen'. Pero, tal vez, esto último no signifique otra cosa que: 'Un río con meandros erosiona su margen, a menos que no la erosione'. Por supuesto, esto es con toda seguridad, predictivamente adecuado. Nada de lo que suceda lo refutará, nada de lo que sucediera podría refutarlo."

Evidentemente, algo ha salido mal. Porque 'En igualdad de circunstancias, un río con meandros erosiona su margen' no es ni falsa ni vacua, y no quiere decir 'Un río con meandros erosiona su margen, a menos que no la erosione'. Supongo que el relato [*story*] de cómo las generalizaciones de las ciencias especiales se las arreglan para estar complementadas y ser informativas a la vez (o, si se prefiere, cómo logran respaldar [*support*] contrafácticos, aunque tengan excepciones), es largo. Formular ese relato es parte de lo que hay que hacer para mostrar por qué razón tenemos ciencias especiales; por qué no tenemos sólo la física básica (véase Fodor, SS). Es también parte de lo que hay que hacer para mostrar cómo funciona la idealización en ciencia. Porque, seguramente,

'*Ceteris paribus*, un río con meandros erosiona sus márgenes' significa algo así como 'Un río con meandros erosiona su margen en cualquier mundo nomológicamente posible en el que las idealizaciones operativas de la geología sean satisfechas'. Que esto es, en general, más fuerte que '*P* en cualquier mundo donde no no-*P*', es indudable. De modo que si, como parece, la psicología de sentido común se apoya en cláusulas *ceteris paribus*, también lo hace la geología.

Hay, entonces, una semejanza ostensiva entre el modo en que las generalizaciones implícitas funcionan en la psicología de sentido común y el modo en que las generalizaciones explícitas funcionan en las ciencias especiales. Pero tal vez esta semejanza sea *meramente* superficial. Donald Davidson es célebre por haber argumentado que las generalizaciones de la ciencia real, a diferencia de las que subyacen a las explicaciones en términos de deseos y creencias de sentido común, son "corregibles". En la ciencia real, y no en las ciencias intencionales, podemos (al menos en principio) eliminar las cláusulas *ceteris paribus* enumerando las condiciones bajo las cuales se supone que las generalizaciones valen.

Según este criterio, sin embargo, la única ciencia real es la física básica. Porque no es verdad que podamos, aun en principio, especificar las condiciones bajo las cuales —digamos— las generalizaciones geológicas valen, *en tanto nos atengamos al vocabulario de la geología*. O para decirlo de un modo menos formal: las causas de las excepciones de las generalizaciones geológicas no son, típicamente, en sí mismas eventos *geológicos*. Pruebe y vea: 'Un río con meandros erosiona su margen a menos que, por ejemplo, el clima cambie y el río se seque'. Pero 'clima' no es un término de la *geología*, ni lo son 'el mundo se termine', 'alguien construya un dique', y un número indefinido de descriptores [*descriptors*] que se requieren para especificar el tipo de cosas que pueden salir mal. Todo lo que uno puede decir con alguna utilidad es: si la generalización dejó de valer, entonces, de alguna manera, las idealizaciones operativas no tienen que haber sido satisfechas. Pero lo mismo sucede en la psicología de sentido común: si no llegó cuando tenía intenciones de llegar, entonces algo debe de haber salido mal.

Las excepciones a las generalizaciones de una ciencia especial son *inexplicables*, típicamente, desde el punto de vista (es decir, en el vocabulario) de dicha ciencia. Esa es una de las cosas que la hace una ciencia *especial*. Pero por supuesto, puede ser posible, sin embargo, explicar las excepciones *en el vocabulario de alguna otra ciencia*. En los casos más familiares, uno 'baja' uno o dos niveles y usa el vocabulario de una cien-

cia más 'básica'. (La corriente dejó de circular por el circuito porque las terminales estaban oxidadas; ya no reconoce objetos familiares porque tuvo un accidente cerebral, y así sucesivamente.) Tener a mano esta estrategia es una de las ventajas que nos proporciona la organización jerárquica de nuestras ciencias. De todos modos —para decirlo brevemente— el mismo esquema que vale para las ciencias especiales, parece valer también para la psicología de sentido común. De un lado, sus cláusulas *ceteris paribus* son ineliminables desde el punto de vista de los recursos conceptuales propios. Pero, del otro lado, no tenemos —al menos por ahora— ninguna razón para dudar de que pueden descargarse en el vocabulario de alguna ciencia de nivel más bajo (digamos, la neurología, o la bioquímica; en el peor de los casos, la física).

Si el mundo es susceptible de ser descripto como un sistema causal cerrado, sólo es descriptible en el vocabulario de nuestra ciencia más básica. De esto no sigue nada que deba preocupar a un psicólogo (o a un geólogo).

Dejo aquí la digresión. La moraleja es que la adecuación predictiva de la psicología de sentido común está más allá de una discusión racional, y no hay tampoco ninguna razón para suponer que se la logra haciendo trampa. Si alguien quiere saber dónde estará mi cuerpo físico el próximo jueves, la mecánica —después de todo, la mejor ciencia de los objetos de tamaño mediano, y con muy buena reputación en su campo— *no nos sirve de nada*. Con mucho, la mejor manera de descubrirlo (habitualmente, en la práctica, la *única* manera de descubrirlo) es: *¡pregúnteme!*

2. *La profundidad de la teoría*

Resulta tentador pensar a la psicología de sentido común como una colección de las perogrulladas que aprendemos en la falda de la abuelita: que el niño que se ha quemado le teme al fuego; que todo el mundo ama a alguien; que el dinero no hace la felicidad; que el refuerzo afecta la tasa de respuesta, y que el camino al corazón de un hombre pasa por su estómago. Coincido en que no vale la pena rescatar ninguna de estas afirmaciones. Sin embargo, como el simple ejemplo esbozado arriba pone en evidencia, subsumir en trivialidades *no* es la forma típica de la explicación psicológica de sentido común. Más bien, cuando tales explicaciones se explicitan, se ve que exhiben a menudo la 'estructura deductiva' característica de la explicación en la ciencia real. Esto tiene dos partes: las

generalizaciones que subyacen a la teoría se definen en términos de inobservables, y conducen a sus predicciones por iteración e interacción más que por ser instanciadas [*instanciated*] de manera directa.

Hermia, por ejemplo, no es una tonta y no es conductista; es perfectamente conciente de que los estados mentales de Demetrio son los que causan su conducta y de que el patrón [*pattern*] de tal causación es típicamente intrincado. En particular, no hay ninguna generalización plausible que respalde contrafácticos [*counterfactual-supporting generalization*] de la forma (*x*) (*y*) (*x es un rival de y*) ⊃ (*x mata a y*). Nada de esto es remotamente verdadero; ni siquiera *ceteris paribus*. Más bien, la generalización que Hermia supone operativa —la que *es* verdadera y puede respaldar contrafácticos— debe ser algo así como *Si x es el rival de y, entonces x prefiere la derrota de y, en igualdad de circunstancias.* Este principio, sin embargo, no menciona conducta; lleva a predicciones de la conducta, pero sólo a través de una serie de otras presuposiciones sobre cómo las preferencias de la gente pueden afectar sus acciones en situaciones dadas. O mejor aún, dado que probablemente no hay generalizaciones que conecten preferencias con acciones prescindiendo de las creencias, en lo que Hermia debe estar confiando es en una teoría implícita de cómo interactúan las creencias, las preferencias y las conductas; nada menos que una teoría implícita de la decisión.

Es un hecho profundo del mundo el que las generalizaciones etiológicas más poderosas valgan respecto de causas inobservables. Hechos tales modelan nuestra ciencia (¡mejor que lo hagan!). Por lo tanto, una prueba de la profundidad de una teoría es que muchas de sus generalizaciones subsuman interacciones entre inobservables. Según este test, cabe presumir que nuestra *meteorología* implícita de sentido común *no* es una teoría profunda, ya que consiste principalmente en generalizaciones prácticas del tipo "Viento del este, lluvia como peste". Concordantemente, el razonamiento que media las aplicaciones de la meteorología de sentido común probablemente no involucre mucho más que la ejemplificación y el *modus ponens*. (Siendo así, no debe sorprender, tal vez, que la meteorología de sentido común no funcione demasiado bien.) La psicología de sentido común, en contraposición, aprueba el examen. Da por sentado que la conducta pública aparece al final de una cadena causal cuyos eslabones son eventos mentales (y por ello inobservables), que puede ser arbitrariamente larga (y arbitrariamente intrincada). Como Hermia, todos nosotros —supongo que literalmente— nacemos mentalistas y realistas, y seguimos siéndolo hasta que el sentido común es expulsado por la mala filosofía.

3. Su indispensabilidad

En la práctica, no tenemos ninguna opción alternativa para el vocabulario de la explicación psicológica de sentido común; no tenemos ningún otro modo de describir nuestras conductas y sus causas, si queremos que nuestras conductas y sus causas sean subsumidas en alguna generalización que conozcamos y que respalde contrafácticos. Nuevamente, esto es difícil de ver porque está demasiado cerca de nosotros.

Por ejemplo, unos párrafos más arriba hablé de la siguiente generalización psicológica de sentido común: *la gente generalmente hace lo que dice que hará*, y dije que cubría el hiato entre un intercambio de proferencias ("Vendrá usted a dar una conferencia...", "Llegaré al aeropuerto el jueves...") y las consiguientes conductas de los hablantes (mi llegada al aeropuerto, el que él esté allí esperándome). Pero esto subestima los argumentos a favor de la indispensabilidad de la psicología de sentido común, dado que sin ella no podemos ni siquiera describir las proferencias como formas lingüísticas (para no hablar de describir las conductas siguientes como tipos de actos). *Palabra* es una categoría *psicológica*. (Es, por cierto, *irreductiblemente* psicológica, hasta donde sabemos; por ejemplo, no hay ninguna propiedad acústica que tengan que compartir todos y solamente los ejemplares [*tokens*] de la misma palabra tipo [*word type*]. Sorprendentemente, no hay en realidad ninguna propiedad acústica que tengan que compartir todos y solamente aquellos ejemplares *enteramente inteligibles* de la misma palabra tipo. Ésta es la razón por la que nuestra mejor tecnología es incapaz, en la actualidad, de construir una máquina de escribir a la que se le pueda dictar.)

Tal como están las cosas —para decirlo claramente— no tenemos *ningún* vocabulario que nos permita especificar tipos de eventos que cumplan con las siguientes cuatro condiciones:

1. Mi conducta al proferir 'Llegaré allí el jueves...' cuenta como un evento del tipo T_i.

2. Mi llegada allí el jueves cuenta como un evento del tipo T_j.

3. 'Los eventos del tipo T_j son consecuencia de eventos del tipo T_i' es aproximadamente verdadera y respalda contrafácticos.

4. Las categorías T_i y T_j no son irreductiblemente psicológicas.

Como las únicas taxonomías conocidas que cumplen con las condiciones 1-3 reconocen tipos de eventos tales como proferir la *forma lingüística* 'Llegaré allí el jueves', o *decir que* uno llegará allí el jueves, o *realizar el acto* de encontrarse con alguien en el aeropuerto, entonces no cumplen con la condición 4.

Los filósofos y los psicólogos solían soñar con un aparato conceptual alternativo, un aparato en el que el inventario de sentido común de tipos de *conducta* sería reemplazado por un inventario de tipos de *movimientos*; las generalizaciones psicológicas que respaldan contrafácticos exhibirían entonces la contingencia de esos movimientos respecto de las variables ambientales y/u orgánicas. No puede negarse, supongo, que la conducta es efectivamente contingente respecto de variables ambientales u orgánicas; sin embargo, las generalizaciones no aparecen. ¿Por qué? Hay una respuesta estándar: es porque la conducta consiste en acciones y la clasificación de las acciones no coincide con la de los movimientos [*actions cross-classify movements*]. La generalización es que el niño que se ha quemado le teme al fuego; pero qué movimiento va a constituir una evitación, depende de dónde esté el niño, dónde esté el fuego..., y así sucesiva y monótonamente. Si uno quiere saber qué generalizaciones subsumen un evento conductual, tiene que saber a qué *acción tipo* pertenece; saber a qué *movimiento tipo* pertenece, por lo común no sirve de nada. Creo en todo esto a pie juntillas.

Sin embargo, por lo general se da por sentado que esta situación *tiene* que ser solucionable, al menos en principio. Después de todo, las generalizaciones de una física completa [*completed physics*] presumiblemente subsumirán todo movimiento de todas las cosas y, por lo tanto, los movimientos de los organismos *inter alia*. Así que si esperamos lo suficiente tendremos, después de todo, generalizaciones capaces de respaldar contrafácticos, que subsumirán los movimientos de los organismos *bajo esa descripción*. Presumiblemente, Dios ya las tiene.

Sin embargo, esto es un tanto engañoso. Ya que las generalizaciones (putativas) de la física completa (putativa) se aplicarían a los movimientos de los organismos *qua* movimientos, pero no *qua* movimientos orgánicos [*organismic*]. Presumiblemente, la física sirve tan poco a las categorías de la macrobiología como a las categorías de la psicología de sentido común. Hace desaparecer tanto al *actor* de la conducta como a la *conducta* misma. Lo que resta son átomos en el vacío. La subsunción de los movimientos de los organismos —y de todo lo demás— en las generalizaciones de la física capaces de respaldar contrafácticos, no garantiza por lo tanto que exista una ciencia cuya ontología reconozca

organismos y sus movimientos. Es decir: la subsumción de los movimientos de los organismos —y de todo lo demás— en las leyes de la física no garantiza que existan leyes acerca de los movimientos de los organismos *qua* movimientos de los organismos. Hasta donde sabemos —excepto tal vez, una pequeña sección de la psicología clásica de los reflejos— no existen tales leyes, y no hay ninguna razón metafísica para suponer que las haya.[1]

Pero todo esto son pamplinas. Aun si la psicología fuera descartable *en principio*, éste no sería un buen argumento para prescindir de ella. (Tal vez la geología sea prescindible en principio; después de todo, todo río es un objeto físico. ¿Sería esa una razón para suponer que los ríos no son una clase natural? ¿O que 'Los ríos con meandros erosionan su margen' es falsa?) Lo relevante con respecto a si vale la pena defender a la psicología de sentido común es su prescindibilidad *de hecho*. Y aquí la situación es completamente clara. No tenemos idea de cómo explicarnos a nosotros mismos si no es en términos de un vocabulario *saturado* de la psicología de creencias/deseos. Uno se siente tentado a formular un argumento trascendental: lo que Kant dijo a Hume acerca de los objetos físicos vale mutatis mutandis para las actitudes proposicionales: no podemos abandonarlas *porque no sabemos cómo hacerlo*.[2]

De tal modo, tal vez sea mejor que tratemos de conservarlas. Conservar las actitudes —reivindicar la psicología de sentido común— significa mostrar cómo podríamos tener (o, como mínimo, mostrar *que* podríamos tener) una ciencia respetable, cuya ontología reconoce explícitamente estados que exhiben la clase de propiedades que el sentido común atribuye a las actitudes. El resto del libro es acerca de esto. Una empresa tal presupone, sin embargo, algún consenso sobre qué tipo de propiedades atribuye realmente el sentido común a las actitudes. Éste es el tema de los próximos párrafos de este capítulo.

1. Quizás haya leyes que relacionan los *estados cerebrales* de los organismos con sus movimientos. Pero, quizá no las haya, puesto que parece posible que las conexiones legales valgan entre los estados del cerebro y las *acciones*, y la clasificación de las acciones, como es usual, no coincida con la de los movimientos. Quizás, esto es lo que se podría predecir al reflexionar. Realmente, ¿podríamos esperar acaso que el mismo estado cerebral que causa la proferencia de 'mono' en los ejemplares [*tokens*] de 'mono', sea el que la causa en los casos de 'monopatín'? ¿Qué decir de las proferencias de (la sucesión fonética) [*empedokliz lipt*] cuando uno habla español y cuando uno habla alemán?

2. El problema con los argumentos trascendentales es, sin embargo, que no es obvio por qué una teoría no podría ser a la vez indispensable y *falsa*. No desearía tener que aceptar una deducción trascendental de las actitudes si el operacionalismo fuera el precio que tengo que pagar por ella.

La esencia de las actitudes

¿Cómo podemos decidir si una psicología *es* una psicología de creencias/deseos? En general, ¿cómo sabemos si las actitudes proposicionales están entre las entidades que reconoce la ontología de una teoría? Estos tipos de preguntas dan lugar a problemas familiares y complicados de identificación interteórica. ¿Cómo distingue uno la eliminación de la reducción y de la reconstrucción? ¿Es correcta la afirmación que no hay tal cosa como la materia deflogistizada, o es 'deflogistización' sólo un término para la oxidación? Aun los conductistas tuvieron problemas en decidir si deseaban negar la existencia de lo mental o afirmar su identidad con lo conductual. (A veces hicieron ambas cosas en oraciones sucesivas. ¡Ah! En esa época sí que sabían vivir sin preocupaciones.)

Propongo hacer la siguiente estipulación. Consideraré que una psicología es de sentido común acerca de las actitudes —de hecho, que las asume— sólo en caso de que postule estados (entidades, eventos o cualquiera otra cosa) que satisfagan las siguientes condiciones:

(i) son semánticamente evaluables;
(ii) tienen poderes causales;
(iii) las generalizaciones implícitas de la psicología de creencias/deseos de sentido común son, en gran parte, verdaderas de tales estados.

Estoy suponiendo, en efecto, que (i)-(iii) son las propiedades esenciales de las actitudes. Ello me parece intuitivamente plausible; si no le parece a usted, dejémoslo ahí. Reñir acerca de intuiciones me resulta vulgar.

Un comentario acerca de estas condiciones.

(i) La evaluación semántica

Las creencias son tipos de cosas que son verdaderas o falsas, los deseos son tipos de cosas que se frustran o satisfacen, los presentimientos [*hunches*] son tipos de cosas que resultan ser correctas o incorrectas, y así sucesivamente. Supondré que lo que hace verdadera (/falsa) a una creencia es algo acerca de su relación con el mundo no psicológico (y no, por ejemplo, algo acerca de su relación con otras creencias, a menos que resulte ser una creencia acerca de creencias). En consecuencia, decir

de una creencia que es verdadera (/falsa) es evaluar esa creencia en términos de su relación con el mundo. Llamaré 'semánticas' a tales evaluaciones. Lo mismo vale, mutatis mutandis, para los deseos, los presentimientos, etcétera.

Como señalé [en el Prefacio], es un enigma acerca de las creencias, los deseos y otros tipos similares, el que sean semánticamente evaluables; casi nada más lo es. (No lo son los árboles, no lo son los números, no lo es la gente. Las proposiciones lo *son* [suponiendo que existan cosas tales], pero eso es poco sorprendente; las proposiciones existen para ser aquello *hacia* [*toward*] lo que son las actitudes: las creencias y deseos.) Veremos más adelante que la evaluabilidad semántica de las creencias/deseos es lo que las convierte, básicamente, en un problema filosófico, y que una defensa de la psicología de creencias/deseos tiene que ser una defensa de tal evaluabilidad.

Algunas veces hablaré del *contenido* de un estado psicológico en lugar de su evaluabilidad semántica. Estas dos ideas están íntimamente interconectadas. Considérese —para cambiar de obra— la creencia de Hamlet de que su tío mató a su padre. Esa creencia tiene un cierto valor semántico; en particular, es una creencia *verdadera*. ¿Por qué verdadera? Bien, porque corresponde a un cierto hecho. ¿Qué hecho? Bien, el hecho de que el tío de Hamlet mató al padre de Hamlet. Pero, ¿por qué es *ese* hecho el que determina la evaluación semántica de la creencia de Hamlet? ¿Por qué no el hecho de que dos es un número primo o el hecho de que Demetrio no mató a Lisandro? Bien, porque el *contenido* de la creencia de Hamlet es *que* su tío mató a su padre (Si se prefiere, la creencia 'expresa la proposición' de que el tío de Hamlet mató a su padre). *Si uno sabe cuál es el contenido de una creencia entonces sabe qué es lo que en el mundo determina la evaluación semántica de la creencia*; así es como se conectan, básicamente, las nociones de contenido y de evaluación semántica.

En este estadio me propongo decir muy poco acerca del contenido; ya llegará el momento. Baste sólo con agregar que las actitudes proposicionales tienen sus contenidos de modo esencial: la manera canónica de identificar una actitud es decir (a) qué tipo de actitud es (una creencia, un deseo, un presentimiento, o cualquiera otra cosa) y (b) cuál es el contenido de la actitud (que el tío de Hamlet mató a su padre, que dos es un número primo, que Hermia cree que Demetrio siente antipatía por Lisandro, o cualquiera otra cosa). En lo que sigue, nada contará como una psicología de actitudes proposicionales —como una reducción, reconstrucción o vindicación de la explicación de creencias/deseos

de sentido común, a menos que reconozca estados que puedan ser individuados de esa manera.

(ii) Los poderes causales

La explicación psicológica de sentido común está profundamente comprometida con al menos tres tipos de causación mental: la causación de la conducta por los estados mentales, la causación de los estados mentales por los eventos del entorno que los afecta (por 'estimulación proximal' ['*proximal stimulation*'], como dicen a veces los psicólogos) y —de alguna manera, las etiologías psicológicas de sentido común más interesantes— la causación de estados mentales entre sí. Como ejemplo de este último tipo, el sentido común reconoce *cadenas de pensamiento* [*chains of thought*] como especies de eventos mentales complejos. Una cadena de pensamiento es presumiblemente una cadena *causal* en la cual un estado mental semánticamente evaluable da lugar a otro; un proceso que a menudo termina en la fijación [*fixation*] de una creencia. (Como se recordará, éste era el tipo de cosa en la que se suponía que Sherlock Holmes era bueno.)

Toda psicología que sea Realista acerca de lo mental reconoce ipso facto sus poderes causales.[3] Los filósofos de la corriente 'funcionalista' sostienen, además, que los poderes causales de un estado mental determinan su identidad (que para que un estado mental sea, por ejemplo, el estado de creer que Demetrio mató a Lisandro, tiene que tener una constelación característica de relaciones causales potenciales y actuales). Ésta es una posición que posee cierto interés para nosotros, puesto que si es verdadera —y si es también verdadero que las actitudes proposicionales tienen sus contenidos de modo esencial— se sigue que los poderes causales de un estado mental determinan de alguna manera su contenido. Sin embargo, no creo que esto sea verdad. Volveremos sobre este punto.

Por ahora, lo importante es esto: lo característico de la psicología de creencias/deseos de sentido común —y en consecuencia de cualquier

3. Negar el compromiso etiológico de los estados mentales fue realmente lo que trató de hacer el conductismo; es lo que 'los conductistas lógicos' y los 'eliminativistas' tuvieron en común. Así por ejemplo, sostener —como Ryle hizo, más o menos— que los estados mentales son especies de disposiciones, es negarse a declarar causales, en un sentido literal, a explicaciones psicológicas tales como "Él lo hizo con la intención de complacerla" o "Su dolor de cabeza le hizo quejarse", para no decir "El mero pensar en dar una conferencia lo hizo enfermar". (Para la discusión véase Fodor, SSA.)

teoría explícita que estoy dispuesto a visualizar como vindicando la psicología de creencias/deseos de sentido común— es el atribuir contenidos y poderes causales a *las mismas cosas mentales que considera semánticamente evaluables*. Es la creencia de Hamlet de que Claudio mató a su padre —exactamente la misma creencia que es verdadera o falsa en virtud de los hechos relativos a la muerte de su padre— lo que causa que se comporte con Gertrudis de una manera tan brutal.[4]

En realidad hay aquí una cuestión más importante. No es sólo que en una psicología de las actitudes proposicionales el contenido y los poderes causales se atribuyen a las mismas cosas. Ocurre también que las relaciones causales entre las actitudes proposicionales logran, de algún modo típico, respetar sus relaciones de contenido, y a menudo las explicaciones en términos de creencias/deseos se basan en eso. Hamlet creía que alguien había matado a su padre porque él creía que Claudio había matado a su padre. El que tuviera su segunda creencia explica que tuviera la primera. ¿Cómo? Bien, presumiblemente *vía* alguna generalización causal tal como 'Si alguien cree *Fa* entonces, *ceteris paribus*,

4. Algunos filósofos imponen aquí una distinción objeto/estado (puede ser objeto/evento), tal que lo que tiene *poderes causales* son las ejemplificaciones de estados mentales tipo (por ejemplo, el *creer* por parte de Hamlet [*Hamlet's believing*], que Claudio mató a su padre), pero lo que tiene *valores semánticos* son las *proposiciones* (por ejemplo, la proposición de que Claudio mató al padre de Hamlet). El punto es que suena extraño decir que el *creer* por parte de Hamlet que *p*, sea verdadero, pero que es correcto decir que la *creencia* de Hamlet (*Hamlet's belief*) de que *p*, lo es.

No estoy convencido de que, a la larga, esa distinción llegue a preocuparme, porque que suene extraño es el menor de mis problemas; a la larga, espero poder pasármelas sin proposiciones. Sin embargo, si usted es escrupuloso con la ontología, me parece bien. En la especie, el texto debería ser: la psicología de deseos y creencias atribuye propiedades causales a las mismas cosas (esto es, a ejemplares de ciertos estados mentales tipo) a las que atribuye objetos proposicionales. Es verdad que el creer, por parte de Hamlet, que Claudio mató a su padre, está implicado en la etiología de su comportamiento hacia Gertrudis [*Gertrudeward*] y tiene como objeto una cierta creencia, esto es, la proposición de que Claudio mató a su padre. Si decimos, entonces, del *estado* de Hamlet de creer que Claudio mató a su padre (o del evento que consiste en la ejemplificación de ese estado), que es semánticamente evaluable, podemos tomarlo como una abreviatura de una manera más precisa de hablar. El estado S tiene el valor semántico V si y sólo si S tiene como su objeto una proposición cuyo valor es V.

Va de suyo que nada de este merodeo ontológico permite el más mínimo avance en lo que hace a la resolución de los enigmas de la intencionalidad. Si (en mi modo de hablar) resulta metafísicamente preocupante que los deseos y las creencias sean semánticamente evaluables, y que los árboles, las rocas y los números naturales no lo sean, es igualmente preocupante que (según la manera ortodoxa de hablar) los creeres [*believings*] tengan objetos proposicionales y los árboles, las rocas y los números primos, no.

cree $\exists x\ (Fx)$'. Esta generalización especifica una relación causal entre dos tipos de estados mentales seleccionados por referencia a (la forma lógica de) las proposiciones que expresan; tenemos así el patrón usual de una atribución simultánea de contenido y de poderes causales. Sin embargo, el punto es que los contenidos de los estados mentales que subsume la generalización causal están ellos mismos semánticamente relacionados; *Fa implica* $\exists x\ (Fx)$, de modo que, por supuesto, el valor semántico de la segunda creencia no es independiente del valor semántico de la primera.

O comparemos el patrón del razonamiento implícito atribuido a Hermia al comienzo de este capítulo. Sugerí que ella tenía que confiar crucialmente en alguna generalización causal como: 'Si *x* quiere que *P*, y *x* cree que no *P* a menos que *Q*, y *x* cree que está en su poder producir *Q* entonces, *ceteris paribus*, *x* trata de producir *Q*'. El sentido común parece sostener muy claramente que algo así es verdadero y tiene respaldo contrafáctico, de ahí que uno explica el intento de *x* de producir *Q*, si uno muestra que *x* tenía las creencias y los deseos de la clase que especifica la generalización. Lo que es absolutamente típico es (a) la apelación a relaciones causales entre estados mentales semánticamente evaluables, como parte integrante de la explicación, y (b) la existencia de relaciones de contenido entre los estados mentales así apelados.

Préstese atención a las letras esquemáticas recurrentes: ellas funcionan, precisamente, para restringir las relaciones de contenido entre los estados mentales que subsume la generalización. Así, en un caso dado, a menos que lo que *x* quiera sea lo mismo que lo que cree que no puede tener sin *Q*, y a menos que lo que *x* cree que se requiere para *P* sea lo mismo que lo que *x* trata de producir, la generalización no se satisface y la explicación fracasa. Es autoevidente que los principios explicativos de la psicología de sentido común logran generalidad mediante la cuantificación sobre agentes (el 'silogismo práctico' pretende aplicarse, *ceteris paribus*, a todos los *x*). Pero, hay que enfatizar que también logran generalidad abstrayendo sobre *contenidos* ('Si usted quiere que *P* y cree que no *P* a menos que *Q*... trata de producir *Q* cualesquiera que sean los *P* y los *Q*'.) Esta última estrategia opera sólo porque muy a menudo los mismos *P* y *Q* —los mismos contenidos— se repiten en estados mentales causalmente relacionados; es decir, sólo porque las relaciones causales respetan con frecuencia las relaciones semánticas.

Este paralelismo entre poderes causales y contenidos engendra lo que es, sin duda, uno de los hechos más notables acerca de la mente cognitiva, tal como la concibe la psicología de creencias/deseos de sentido

común: la frecuente similaridad entre series de pensamientos y *argumentos*. Aquí tenemos a Sherlock Holmes, al final de "The Speckled Band", haciendo lo que sabe hacer:

> Reconsideré mi posición al instante cuando... me di cuenta de que cualquier peligro que amenazara a un ocupante de la habitación no podía venir ni de la ventana ni de la puerta. Rápidamente mi atención se dirigió, como ya se lo señalé, a este ventilador y al llamador que colgaba encima de la cama. El descubrimiento de que era de utilería y de que la cama estaba atornillada al piso, inmediatamente me hizo sospechar que la cuerda estaba allí como un puente para que algo pasara a través del agujero y llegara a la cama. Inmediatamente se me ocurrió la idea de una víbora, y cuando la asocié con la información de que el Doctor había traído de la India un conjunto de animales, tuve la sensación de que probablemente estaba en el camino correcto.

El pasaje pretende ser un trozo de psicología reconstructiva: un fragmento histórico de la sucesión de estados mentales que llevaron a Holmes primero a sospechar y luego a creer que el doctor hizo lo que hizo con su víbora. Lo que es interesante aquí para nuestros propósitos es que el relato de Holmes no es *solamente* psicología reconstructiva. Realiza una doble función, ya que sirve también para reunir las *premisas* con la *conclusión*, en una inferencia plausible de que el doctor hizo lo que hizo con la víbora. Debido a que su secuencia de pensamientos es similar a un argumento, Holmes espera que Watson se *convenza* en virtud de las consideraciones que, cuando se le ocurrieron a Holmes, causaron su propia convicción. Lo que conecta el aspecto histórico-causal del relato de Holmes con la apariencia de ser una inferencia-plausible, es el hecho de que los pensamientos que fijan la creencia de que *P* proporcionan, bastante a menudo, *fundamentos* razonables para creer que *P*. Si no fuera éste el caso —si no hubiera esta armonía general entre las propiedades semánticas y causales de los pensamientos, tal que, como Holmes dice en otra historia, "una inferencia verdadera invariablemente sugiere otras"— pensar no tendría, después de todo, mucha utilidad.

Todo esto sugiere un conjunto de cuestiones filosóficas: exactamente, ¿*qué* clases de relaciones de contenido están preservadas en las generalizaciones que subsumen los casos típicos de causación de creencias/deseos? Y —en muchos sentidos una cuestión más difícil— ¿cómo puede la mente estar construida de manera tal que resulten verdaderas de ella dichas generalizaciones? ¿Qué tipo de mecanismo podría tener estados que estén conectados tanto semántica como causalmente, tal

que las conexiones causales respeten las conexiones semánticas? Es el carácter espinoso de esas cuestiones lo que causa que muchos filósofos hayan perdido las esperanzas respecto de la psicología de sentido común. Pero, por supuesto, el argumento tiene doble filo: si el paralelismo entre contenido y relaciones causales es, como parece ser, un rasgo profundo de la mente cognitiva, entonces, a menos que podamos salvar la noción de contenido, hay un rasgo profundo de la mente cognitiva que nuestra psicología va a pasar por alto.

(iii) La preservación de las generalizaciones

Lo que he dicho hasta aquí equivale en gran medida a esto: una psicología explícita que reivindique las explicaciones de creencias/deseos de sentido común, tiene que permitir la asignación de contenido a los estados mentales causalmente eficaces y tiene que reconocer las explicaciones conductuales en las cuales las generalizaciones abarcantes se refieren a (o cuantifican sobre) los contenidos de los estados mentales que subsumen. Ahora agrego que las generalizaciones que son reconocidas por la teoría reivindicativa no tienen que ser *disparatadas* desde el punto de vista del sentido común; los poderes causales de las actitudes tienen que ser, más o menos, lo que el sentido común supone que son. Después de todo, la psicología de sentido común no será reivindicada a no ser que resulte verdadera, al menos, aproximadamente.

Sin embargo, carezco de un catálogo de generalizaciones de sentido común que tengan que ser reconocidas por una teoría, si quiere comprometerse ontológicamente con actitudes proposicionales genuinas. Mucho de lo que cree el sentido común acerca de las actitudes seguramente tiene que ser falso (mucho de lo que el sentido común cree acerca de *cualquier cosa* seguramente tiene que ser falso). Uno espera, más bien, que haya muchas más cosas en la mente —y mucho más extrañas— que lo que el sentido común ha soñado; sino, ¿qué gracia tendría hacer psicología? Los indicios son, y lo han sido desde Freud, que esta esperanza será ampliamente satisfecha. Por ejemplo, en contra del sentido común, parece que mucho de lo que hay en la mente es inconciente y en contra del sentido común, parece que mucho de lo que hay en la mente no es aprendido. No me perturbo, conservo la serenidad.

Por lo demás, hay una gran parte de la psicología de sentido común de la que no tenemos razones para dudar —por lo menos hasta ahora— y que los partidarios de las actitudes odiarían abandonar. Por lo tanto, es difícil imaginar una psicología de la acción que esté compro-

metida con las actitudes pero que no reconozca algunas de las relaciones causales entre creencias, deseos e intenciones conductuales (las "máximas" de los actos) que explican las teorías de la decisión. De modo similar, es difícil imaginar una psicolingüística (para el castellano) que atribuya creencias, deseos, intenciones comunicativas, etcétera a hablantes/oyentes, pero que no implique una infinidad de teoremas similares a éstos:

- 'Demetrio mató a Lisandro' es la forma lingüística estándar usada para comunicar la creencia de que Demetrio mató a Lisandro.

- 'El gato está sobre el felpudo' es la forma lingüística estándar usada para comunicar la creencia de que el gato está sobre el felpudo.

- 'Demetrio mató a Lisandro o el gato está sobre el felpudo' es la forma lingüística estándar usada para comunicar la creencia de que Demetrio mató a Lisandro o el gato está sobre el felpudo.

Y así indefinidamente. Por cierto que es difícil imaginar una psicolingüística que apele a las actitudes proposicionales de hablantes/oyentes del castellano para explicar su conducta verbal pero que no implique que ellos *conocen* al menos uno de esos teoremas para cada oración de su lenguaje. De modo que queda por reivindicar una enorme cantidad de sentido común para la psicología.

La moda filosófica hoy en día es un esencialismo seguro de sí mismo. Hay gente que tiene Puntos de Vista Muy Fuertes (a esos puntos de vista se los llama 'intuiciones modales') acerca de si podría haber o no gatos en un mundo en el cual todos los felinos domésticos fueran robots marcianos, y si podría existir o no Homero en un mundo en el que nadie hubiese escrito la *Odisea* o la *Ilíada*, ¡felices de ellos! Su condición epistémica es envidiable, pero yo no aspiro a ella. No sé exactamente cuánta psicología de sentido común tendría que ser verdadera para que haya deseos y creencias. Digamos que al menos un poco, y preferiblemente mucho. Porque no dudo de que buena parte de la psicología de sentido común *es* verdadera, éste es un tema que no me desvela.

TRM

La tesis principal de este libro puede expresarse como sigue: *no tenemos razón para dudar —en realidad, tenemos razones sustanciales*

para creer— que es posible tener una psicología científica que reivindique la explicación de creencias/deseos de sentido común. Pero aunque ésta es mi tesis, no me propongo defenderla en un plano tan abstracto. Porque ya hay en esta área una teoría (más o menos) empírica que, desde mi punto de vista, es interpretable, razonablemente, como estando ontológicamente comprometida con las actitudes, y que —nuevamente, desde mi punto de vista— es muy probable que sea aproximadamente verdadera. Si estoy en lo correcto acerca de esa teoría, ella *es* una reivindicación de las actitudes. Más aún, puesto que es la única cosa de este tipo en el ambiente (es la *única* propuesta para una psicología científica de creencias/deseos que hay en el área), defender los supuestos de sentido común acerca de las actitudes y defender esa teoría resulta ser la misma empresa, extensionalmente hablando.

Sea como fuere, ésta es la estrategia que seguiré: argüiré que las clases de objeciones que los filósofos han esgrimido recientemente en contra de la explicación de creencias/deseos, no son (para decirlo suavemente) concluyentes respecto de la mejor teoría reivindicatoria de que disponemos. Por lo tanto, el resto de este capítulo estará dedicado a esbozar cómo esa teoría trata a las actitudes y por qué un enfoque tal de las actitudes parece tan prometedor. Puesto que este relato es bastante bien conocido, tanto en círculos filosóficos como psicológicos, me propongo ser breve.

Lo que estoy vendiendo es la Teoría Representacional de la Mente (en adelante TRM; para su discusión véase entre otras fuentes: Fodor, PA; Fodor, LOT; Field, MR). En el corazón de la teoría se encuentra la postulación de un lenguaje del pensamiento: un conjunto infinito de 'representaciones mentales' que funcionan a la vez como los objetos inmediatos de las actitudes proposicionales y como los dominios de los procesos mentales. Más precisamente, la TRM es la conjunción de las dos afirmaciones que siguen:

> *Afirmación 1* (la naturaleza de las actitudes proposicionales):
>
> Para todo organismo *O*, y para toda actitud *A* hacia la proposición *P*, existe una relación *R* ('computacional'/'funcional') y una representación mental *RM* tal que
>
> *RM* significa que *P*, y
> *O* tiene *A* si y sólo si *O* tiene *R* con *RM*.

(Luego veremos que el bicondicional necesita ser mitigado un poco, pero no de manera que afecte en demasía el espíritu de la propuesta.)

Se trata de un trazo delgado entre la claridad y la ampulosidad. Un modo más tosco pero más inteligible de formular la afirmación 1, sería éste: creer que tal y cual es tener ejemplificado [*tokened*] en la cabeza, de una cierta manera, un símbolo mental que significa que tal y cual; es tener tal ejemplar 'en la caja de creencias' como diré a veces. Concordantemente, esperar que tal y cual es tener un ejemplar de ese mismo símbolo mental ejemplificado en la cabeza, pero de una manera muy diferente; es tenerlo ejemplificado 'en la caja de esperas'. (La diferencia entre tener el ejemplar en una caja o en otra corresponde a la diferencia entre los roles causales de creencias y deseos. La idea de hablar de cajas de creencias, y similares, como una manera abreviada de representar las actitudes en tanto estados *funcionales*, se debe a Steve Schiffer.) Y así sucesivamente para cada actitud que uno pueda tener hacia una proposición, y así sucesivamente para cada proposición respecto de la cual uno pueda tener una actitud.

> *Afirmación 2* (la naturaleza de los procesos mentales):
> Los procesos mentales son secuencias causales de ejemplificaciones de representaciones mentales.

Una sucesión [*train*] de pensamientos es, por ejemplo, una secuencia causal de ejemplificaciones de representaciones mentales que expresan las proposiciones que son los objetos de los pensamientos. En una primera aproximación, pensar 'Va a llover, me iré adentro'es tener una ejemplificación que significa *me iré adentro*, causada de una cierta manera por una ejemplificación de una representación mental que significa *va a llover*.

Hasta aquí la formulación de TRM.

Pienso que hay varias razones para creer que TRM puede ser más o menos verdadera. La mejor razón es que una u otra versión de TRM subyace prácticamente a toda la investigación psicológica actual sobre lo mental, y nuestra mejor ciencia es ipso facto nuestra mejor aproximación acerca de lo que hay y de qué está hecho lo que hay. Algunos de mis colegas filosóficos no encuentran persuasiva esta clase de argumentos. Me hacen sonrojar. (Para una discusión extensa de cómo la TRM modela el trabajo actual sobre la cognición, véase Fodor, LOT, especialmente el capítulo 1. Para una discusión de la conexión entre la TRM y el Realismo Intencional de sentido común —y algunos argumentos según los cuales, dado el segundo, la primera es prácticamente obligatoria—, véase el Apéndice [de Fodor, P].)

Pero tenemos una razón para conjeturar que la TRM pueda ser ver-

dadera, aun dejando a un lado los detalles de su éxito empírico. Señalé más arriba que hay un paralelismo notable entre las relaciones causales que valen entre los estados mentales, por una parte, y las relaciones semánticas que valen entre sus objetos proposicionales, por la otra; y que las propiedades muy profundas de lo mental —como por ejemplo, que las secuencias de pensamientos preservan en gran medida la verdad— giran en torno a esa simetría. La TRM sugiere un mecanismo plausible para tal relación, y esto es algo que ninguna explicación previa de lo mental ha sido capaz de hacer. Explicitaré esto un poco; ello ayudará a aclarar exactamente *por qué* la TRM ocupa un lugar tan central en el modo como los psicólogos piensan ahora acerca de la mente.

La estratagema consiste en combinar la postulación de representaciones mentales con la 'metáfora del computador'. Los computadores nos muestran cómo conectar las propiedades semánticas con las propiedades causales respecto de los *símbolos*. Así, si tener una actitud proposicional involucra ejemplificar un símbolo, podemos obtener alguna ventaja conectando propiedades semánticas con propiedades causales respecto de los *pensamientos*. En este tema, creo que realmente ha habido algo parecido a un progreso intelectual. Dejando a un lado los detalles técnicos, éste es —desde mi punto de vista— el único aspecto de la ciencia cognitiva contemporánea que representa un avance importante en cuanto a las versiones del mentalismo que la precedieron en los siglos dieciocho y diecinueve. Lo que estaba equivocado en el Asociacionismo, por ejemplo, fue que no había manera de obtener una vida mental *racional* que emergiera de los tipos de relaciones causales entre pensamientos reconocidas por las 'leyes de la asociación'. (Véase en las páginas finales del *Ulises* de Joyce una parodia —presumiblemente inadvertida— del punto de vista contrario.)

He aquí, en dos palabras, cómo se supone que sigue el nuevo relato: uno conecta las propiedades causales de un símbolo con sus propiedades semánticas *via su sintaxis*. La sintaxis de un símbolo es una de sus propiedades físicas de orden superior. En una primera aproximación metafórica, podemos pensar la estructura sintáctica de un símbolo como un rasgo abstracto de su configuración [*shape*].[5] Debido a que, en realidad,

5. Sin embargo, *cualquier* propiedad nómica de los símbolos ejemplares [*tokens*] —cualquier propiedad en virtud de cuya posesión satisfacen leyes causales— en principio también lo haría. (Así, por ejemplo, la estructura sintáctica podría estar realizada por relaciones entre estados electromagnéticos más que por relaciones entre configuraciones; como es el caso, ciertamente, en las computadoras reales.) Éste es el punto de la doctrina funcionalista de que, en principio, uno puede hacer una mente a partir de casi cualquier cosa.

la sintaxis se reduce a la configuración y a que la configuración de un símbolo es un determinante potencial de su rol causal, es muy fácil ver cómo podrían existir entornos [*environments*] en los cuales el rol causal de un símbolo se correlacionara con su sintaxis. Es decir, es fácil imaginar ejemplificaciones de símbolos que interactúan causalmente en *virtud de* sus estructuras sintácticas. La sintaxis de un símbolo podría determinar las causas y los efectos de sus ejemplificaciones, de la misma manera que la geometría de una llave determina qué cerraduras abrirá.

Pero ahora sabemos, a partir de la lógica moderna, que algunas de las relaciones semánticas entre símbolos pueden ser, por así decir, 'remedadas' ['*mimicked*'] por sus relaciones sintácticas; de esto es de lo que trata una teoría de la prueba, vista a gran distancia. Por lo tanto, dentro de ciertos límites famosos, la relación semántica que vale entre dos símbolos, cuando la proposición expresada por uno es implicada por la proposición expresada por el otro, puede ser remedada por relaciones sintácticas en virtud de las cuales uno de los símbolos es derivable del otro. En consecuencia, podemos construir máquinas que tengan, nuevamente dentro de límites famosos, la siguiente propiedad:

> Las operaciones de la máquina consisten enteramente en transformaciones de símbolos;
>
> en el curso de la realización de tales operaciones la máquina es sensible solamente a las propiedades sintácticas de los símbolos;
>
> y las operaciones que la máquina realiza con los símbolos se limitan, enteramente, a alterar sus configuraciones.

Con todo, la máquina está diseñada de tal modo que transformará un símbolo en otro si y sólo si las proposiciones expresadas por los símbolos así transformados están en ciertas relaciones *semánticas*; por ejemplo, la relación que las premisas tienen con la conclusión en un argumento válido. Tales máquinas —computadores, por supuesto— *son* sólo entornos en los cuales la sintaxis de un símbolo determina su rol causal de modo de respetar su contenido. Ésta es, pienso, una idea realmente fabulosa, como es fabuloso que funcione.

Espero que sea claro cómo se supone que esto se conecta con la TRM y con el compromiso ontológico respecto de las representaciones mentales. Los computadores son una solución al problema de la mediación entre las propiedades causales de los símbolos y sus propiedades

semánticas. Así *si* la mente es una especie de computador, comenzamos a ver cómo se puede tener una teoría de los procesos mentales que tenga éxito, cuando —literalmente— todos los intentos anteriores fracasaron lastimosamente; una teoría que explica cómo podría haber relaciones de contenido no arbitrarias entre pensamientos causalmente relacionados. Pero, evidentemente, para que esta propuesta funcione tiene que haber representaciones mentales. En el diseño computacional, el rol causal es apareado [*is brought into phase with*] con el contenido, aprovechando los paralelismos entre la sintaxis de un símbolo y su semántica. Pero esa idea no le sirve de nada a la teoría de la *mente*, a menos que haya símbolos *mentales*: particulares mentales que posean tanto propiedades sintácticas como semánticas. Tiene que haber símbolos mentales porque, en resumidas cuentas, sólo los símbolos tienen sintaxis, y nuestra mejor teoría disponible de los procesos mentales —en realidad la *única* teoría disponible de los procesos mentales que no se *sabe* que sea falsa— necesita figurar a la mente como una máquina que opera sintácticamente.

A veces, quienes la admiran menos que yo han alegado en contra de la psicología de creencias/deseos de sentido común, que es una teoría "estéril" (véase especialmente Churchland, EMPA; Stich, FFPCS) que es dudoso que haya progresado mucho desde Homero y que no ha progresado nada desde Jane Austin. Indudablemente que hay un sentido en el que esta acusación está justificada; la psicología de sentido común puede ser ciencia implícita pero no es, en la versión de nadie, una ciencia *investigativa* implícita. (Lo que los novelistas y poetas hacen no cuenta como investigación según los criterios austeros actuales.) En suma, si se desea evaluar el progreso se necesita examinar no la teoría implícita de sentido común, sino el mejor candidato para su reivindicación explícita. Y aquí el progreso ha sido enorme. No es sólo que ahora sabemos algo sobre la memoria y la percepción (*qua* medios para la fijación de la creencia) y un poco acerca del lenguaje (*qua* medio para la comunicación de la creencia); véase cualquier libro estándar de psicología. El logro efectivo es que estamos (quizá) a punto de resolver un gran misterio acerca de la mente: *¿cómo podrían sus procesos causales ser semánticamente coherentes*? O si se quiere, con bombos y platillos: *¿cómo es posible, mecánicamente, la racionalidad*?[6] Nótese que este tipo

6. Lo cual no significa negar que haya (¡ejem!) ciertas dificultades técnicas residuales. (Véase, por ejemplo, la parte 4 de Fodor, MOM.) Una teoría de la racionalidad (esto es una teoría de *nuestra* racionalidad) tiene que dar cuenta, no meramente de la 'coherencia

de problema no puede ser enunciado, y mucho menos resuelto, a menos que supongamos —tal como la psicología de creencias/deseos de sentido común requiere— que haya estados mentales con contenidos semánticos y roles causales a la vez. Una buena teoría es una teoría que conduce a preguntas que tienen respuestas, y viceversa, *ceteris paribus*.

Con todo, la TRM no funcionará según la forma tosca expuesta más arriba. Propongo terminar este capítulo refinándola un poco.

De acuerdo con la afirmación 1, la TRM requiere las dos afirmaciones siguientes:

> para cada ejemplificación de una actitud proposicional, existe una ejemplificación de una relación correspondiente entre un organismo y una representación mental;
>
> y
>
> para cada ejemplificación de esa relación, existe una ejemplificación correspondiente de una actitud proposicional.[7]

Esto es, sin embargo, demasiado fuerte: la equivalencia fracasa en ambas direcciones.

Y es, por supuesto, lo que debíamos esperar dada nuestra experiencia en otros casos en los que la ciencia explícita recoge el aparato conceptual de sentido común. Por ejemplo, como todos señalan, sencillamente no es verdad que la química identifique cada muestra de agua con una muestra de H_2O; al menos no lo hace si la noción operativa de agua es la de sentido común, según la cual agua es aquello que bebemos, aquello en lo cual navegamos y con lo cual llenamos nuestras bañeras. Lo que la química hace es reconstruir las categorías de sentido común *en aquello que la teoría misma identifica como casos centrales: el agua químicamente pura* es H_2O. Por supuesto, la infrecuencia ecológica de tales casos centrales no es un argumento en contra de la afirmación de que la ciencia química reivindica la taxonomía del sentido común: el sentido común estaba en lo cierto acerca de que hay tal sustancia como

semántica' de los procesos de pensamiento en abstracto, sino de nuestra habilidad para separar los tipos mismos de inferencias racionales que hacemos. (Tiene que dar cuenta, por ejemplo, de nuestra habilidad para hacer ciencia.) Una teoría tal no estaría disponible la próxima semana.

7. Porque no quiero preocuparme por la ontología de la mente, he evitado enunciar la TRM como una tesis de la identidad. Pero usted puede hacerlo si quiere.

el agua, estaba en lo cierto acerca de que hay agua en el río Nilo y nuevamente estaba en lo cierto acerca de que el agua es aquello que bebemos para apagar nuestra sed. Nunca se dijo que el agua del Nilo era químicamente pura; 'químicamente pura' no es una frase del vocabulario de sentido común.

De manera exactamente similar, la TRM reivindica a la psicología de sentido común en lo que la TRM identifica como casos centrales; en esos casos, lo que el sentido común considera que son ejemplificaciones de actitudes proposicionales resultan ser realmente ejemplificaciones de una relación entre un organismo y una representación mental. A los otros casos —donde tenemos ejemplificaciones de actitudes sin la relación o ejemplificaciones de relaciones sin las actitudes— la teoría los trata como derivativos. Repito, esto es *exactamente* lo que esperaríamos de los antecedentes científicos. Sin embargo, los filósofos han hecho una horrible alharaca respecto de esto al discutir la reivindicación de las actitudes (véase la controversia sobre la 'representación explícita' —o no— de las gramáticas, recientemente protagonizada, entre otros, por Stabler, HAGR y Demopoulos y Matthews, HGMR). De modo que, por un momento, consideremos los detalles. Hacerlo permitirá refinar la afirmación 1, que es lo que queremos.

Caso 1. Actitudes sin representaciones mentales

He aquí un caso tomado de Dennett:

> En una reciente conversación con el diseñador de un programa para jugar al ajedrez, escuché la siguiente crítica a un programa rival: "El programa piensa que debería mover antes su reina". Esto adscribe una actitud proposicional al programa, de un modo muy útil y predictivo, porque como el diseñador dijo, habitualmente uno piensa en perseguir a la reina a lo largo del tablero. Pero, a pesar de los niveles de representación explícita que se encuentran en ese programa, en ninguna parte aparece algo que sea aproximadamente sinónimo de "Yo debería mover antes mi reina". El nivel de análisis al que pertenece el comentario del diseñador describe, de un modo completamente inocente, rasgos del programa que son propiedades emergentes de los procesos computacionales que tienen "realidad ingenieril". No hay razón para creer que la relación entre el discurso acerca de las creencias y el discurso acerca de los procesos psicológicos, sea más directa (CCC, 107; véase también Matthews, TWR).

Nótese que el problema que Dennett plantea no es precisamente el de que algunas de las cosas que el sentido común considera actitudes proposicionales de alguien, sean *disposicionales*. No es como el problema de que yo podría decir ahora que creo alguna consecuencia abstrusa de la teoría de los números —una consecuencia en la que, hablando en términos de sentido común, nunca pensé— porque yo *aceptaría* la prueba del teorema *si* me la mostraran. Por supuesto que es verdad que las creencias meramente disposicionales no podrían corresponder a las ejemplificaciones *efectivas* [*occurrent*] de relaciones con representaciones mentales; por lo tanto, la afirmación 1 tiene que ser reformulada. Pero el problema es superficial puesto que la revisión relevante de la afirmación 1 sería muy obvia; a saber, que para cada creencia *efectiva* hay una ejemplificación *efectiva* de una representación mental que le corresponde, y para cada creencia disposicional hay una *disposición* a ejemplificar una representación mental que le corresponde.

Esto deja abierta una cuestión que se suscita con independencia de los puntos de vista que se puedan tener acerca de la TRM: ¿cuándo son *verdaderas* las atribuciones de creencias disposicionales? Supongo que las creencias disposicionales de uno podrían ser identificadas razonablemente con la clausura [*closure*] de las creencias efectivas de uno bajo principios de inferencia que uno acepta de manera explícita. Y la TRM podría convivir con la vaguedad que resulta al determinar cuáles de las creencias pertenecen a tal clausura. *Qua disposicionales*, las actitudes no desempeñan ningún rol causal en los procesos mentales *reales*; sólo las actitudes efectivas —si vamos al caso sólo lo que es efectivo— son causas reales. Por lo tanto, la TRM puede permitirse ser un tanto operacionalista acerca de las creencias meramente disposicionales (véase Lycan, TB), en la medida en que adopte una línea dura respecto de las creencias efectivas.

Sin embargo, para repetirlo nuevamente, el problema que se plantea en el texto de Dennett, no es de este tipo. No es que el programa crea *potencialmente*, 'Mueva antes su reina'. El punto de Dennett es que el programa opera realmente según este principio, pero no en virtud de una ejemplificación de algún símbolo que lo exprese. Y, por supuesto, el ajedrez no es el único caso. El compromiso conductual con el *modus ponens* o con la regla sintáctica del inglés para formular preguntas ['*wh*'-*movement*] *podría* presagiar que esas reglas [esos ajustes] están inscriptas en escritura cerebral [*brain writing*]. Pero ello no es necesario, puesto que esas reglas podrían ser obedecidas pero no literalmente seguidas, como los filósofos dicen a veces.

En el ejemplo de Dennett, se tiene una actitud que es, por decirlo así, un emergente a partir de su propia implementación. Podría parecer que este modo de expresarse sugiere una manera de salvar la afirmación 1: la máquina no representa explícitamente 'Mueva antes su reina' pero, al menos, podemos suponer que *sí* representa explícitamente algunas reglas del juego más detalladas (las que Dennett dice que tienen "realidad ingenieril"). Para estas reglas, al menos, se satisfaría así una forma fuerte de la afirmación 1. Pero, esta sugerencia tampoco funciona. *Ninguno* de los principios de acuerdo con los cuales opera un sistema computacional precisa estar explícitamente representado por una fórmula ejemplificada en el dispositivo; no hay garantía de que el programa de una máquina esté explícitamente representado en la máquina de la que es programa. (Véase Cummins, IMM; a grandes rasgos, el punto es que para toda máquina que computa una función ejecutando un algoritmo explícito, existe otra máquina —una máquina con un *hardware* característico ['*hardwired*']— que computa la misma función pero *sin* ejecutar un algoritmo explícito.) Por lo tanto, se podría inquirir: después de todo, ¿qué obtiene la TRM de la 'metáfora del computador'?

Incluso hay aquí una cuestión de principio, una cuestión que a veces se lee en el diálogo entre Aquiles y la Tortuga (Lewis Carroll). No todas las reglas de inferencia que un sistema computacional opera *pueden* estar representadas de modo *explícito* en el sistema; algunas de ellas, se dice, tienen que estar 'realizadas en el *hardware*'. De otro modo, la máquina no operaría. Un computador en el que los principios de operación *sólo* están explícitamente representados, es como un pizarrón en el que han sido escritos los principios. Tiene el problema de Hamlet: cuando se pone en funcionamiento, nada ocurre.

Puesto que todo esto es claramente correcto y, puede argüirse, importante, la cuestión que surge es cómo establecer la TRM de manera tal que los casos en los que los programas están realizados de modo característico en el *hardware* [*hardwired*] no cuenten como disconfirmaciones de la afirmación 1. Volveré a esto en un momento. Consideremos primero:

Caso 2. Representaciones mentales sin actitudes

Lo que la TRM toma prestado de los computadores es, en primera instancia, la receta para mecanizar la racionalidad: para explotar los paralelismos entre las propiedades semánticas y las propiedades sintác-

ticas de los símbolos úsese una máquina que opere sintácticamente. Algunas —pero no todas— las versiones de la TRM toman más que esto; no sólo una teoría de la racionalidad, sino también una teoría de la inteligencia. De acuerdo con este relato, la conducta inteligente explota de manera típica una 'arquitectura cognitiva' constituida por *jerarquías* de procesadores simbólicos [*symbol processors*]. En la cúspide de tal jerarquía podría existir una capacidad muy compleja: resolver un problema, formular un plan, emitir una oración. En la base, sin embargo, están sólo los tipos de operaciones no inteligentes que pueden realizar las máquinas de Turing: borrar símbolos, almacenar símbolos, copiar símbolos, etcétera. Llenar los niveles intermedios equivale a reducir —analizar— una capacidad inteligente a un complejo de capacidades tontas; por lo tanto, a una explicación del primer tipo.

He aquí un ejemplo típico de un tipo de teoría representacional que sigue esas líneas:

> Éste es el modo en que atamos nuestros zapatos: hay un hombrecito que vive en nuestra cabeza. El hombrecito tiene una biblioteca. Cuando uno actúa con la intención de atarse los zapatos, el hombrecito busca un volumen titulado *Atarse los zapatos*. El volumen dice cosas tales como: "Tome con la mano izquierda el extremo libre izquierdo del cordón del zapato. Cruce el extremo libre izquierdo del cordón sobre el extremo libre derecho del cordón...", etc... Cuando el hombrecito lee "Tome con la mano izquierda el extremo libre izquierdo del cordón", nos lo imaginamos llamando por teléfono al capataz del taller encargado de tomar cordones. El capataz del taller encara la supervisión de tal actividad de una manera que es, en esencia, un microcosmos de cómo atar el propio zapato. Podría imaginarse al capataz dirigiendo una cuadrilla de esclavos asalariados cuyas funciones incluyen: buscar entre las representaciones de *inputs* visuales huellas de cordones, despachar órdenes para flexionar y contraer dedos de la mano izquierda, etc. (Fodor, ATK, 63-65, ligeramente revisado).

En la cúspide están los estados que bien pueden corresponder a las actitudes proposicionales que el sentido común está dispuesto a reconocer (saber cómo atarse los zapatos, pensar acerca de atar zapatos). Pero en la base y en los niveles intermedios debe de haber, seguramente, una cantidad de operaciones de procesamiento de símbolos que no corresponden a nada que la *gente* hace, por oposición a sus sistemas nerviosos. Éstas son las operaciones de lo que Dennett ha llamado sistemas computacionales "subpersonales"; y aunque satisfacen la actual formulación de la afirmación 1 (en el sentido de que involucran causalmente

ejemplificaciones eficaces de las representaciones mentales), sin embargo no es claro que correspondan a algo que el sentido común consideraría como la ejemplificación de una actitud. Pero, entonces, ¿cómo tenemos que formular la afirmación 1 para evitar su disconfirmación por parte de los procesos de información subpersonal?

La reivindicación vindicada

Hay un sentido en el que estos tipos de objeciones a la afirmación 1 no me parecen que sean muy serios. Como señalé más arriba, la reivindicación por parte de la TRM de la explicación de creencias/deseos, *no* requiere que todo caso que el sentido común considera como la ejemplificación de una actitud deba corresponder a la ejemplificación de una representación mental, o viceversa. Todo lo que requiere es que tales correspondencias rijan en aquello que la teoría reivindicadora toma como casos centrales. Por otra parte, la TRM ha tenido que poder decir qué casos son centrales. La química puede sostener que el río Nilo es irrelevante, en gran medida, para confirmar 'El agua es H_2O', pero sólo porque proporciona fundamentos independientes para negar que lo que está en el Nilo sea una muestra químicamente pura, ¡de cualquier cosa!

Entonces, ¿cuáles son los casos centrales para la TRM? La respuesta debería ser clara a partir de la afirmación 2. De acuerdo con ella los procesos mentales son secuencias causales de transformaciones de representaciones mentales. Se sigue que las ejemplificaciones de actitudes *tienen* que corresponder a ejemplificaciones de representaciones mentales cuando ellas —las ejemplificaciones de actitudes— son episodios [*episodes*] en los procesos mentales. Si los objetos intencionales de tales ejemplificaciones de actitudes causalmente eficaces *no* son explícitamente representados, entonces la TRM es falsa. Repito para enfatizar: si la ocurrencia de un pensamiento es un episodio en un proceso mental, entonces la TRM está comprometida con la representación explícita de su contenido. El lema es, por lo tanto, No hay Causación Intencional sin Representación Explícita.

Nótese que esta manera de elegir los casos centrales es consistente con los contraejemplos alegados. La TRM dice que los contenidos de una secuencia de actitudes que constituye un proceso mental, tienen que ser expresados mediante las ejemplificaciones explícitas de representaciones mentales. Pero las reglas que determinan el curso de la transfor-

mación de esas representaciones —el *modus ponens*, la regla del inglés para la formulación de preguntas, 'Mueva su reina antes' o cualquiera otra cosa— no necesitan ser explícitas. Pueden emerger de los procedimientos de implementación explícitamente representados, o de estructuras del *hardware*, o de ambos. A grandes rasgos: de acuerdo con la TRM, los programas —que corresponden a las 'leyes del pensamiento'— *pueden* ser explícitamente representados, pero las 'estructuras de datos' ['*data structure*'] —que corresponden a los contenidos de los pensamientos— *tienen que serlo*.

De tal modo, en el ejemplo de Dennett acerca del ajedrez, la regla 'Mueva antes la reina' puede o no estar expresada mediante un símbolo 'mental' (/ del lenguaje de programa). Eso depende de cómo opera la máquina; específicamente, depende de si *consultar* la regla es un paso en las operaciones de la máquina. Considero que en la máquina que Dennett tiene en mente, no es un paso; *tener el pensamiento 'Mejor mueva antes la reina', no constituye un episodio en la vida mental de esa máquina*.[8] Pero entonces, el contenido intencional de ese pensamiento *no* necesita estar explícitamente representado para estar en consonancia con que 'no hay causación intencional sin representación explícita' sea verdadero. Por oposición, las representaciones del tablero —los estados posibles o reales del juego— sobre los cuales se definen las computaciones de la máquina *tienen* que ser explícitos, precisamente *porque* las computaciones de la máquina *se definen* sobre las representaciones. Esas computaciones constituyen los 'procesos mentales' de la máquina, por lo tanto, o son sucesiones causales de representaciones explícitas o la teoría representacional del juego de ajedrez es falsa respecto de la máquina. Para decirlo brevemente: restringir la atención al *status* de las reglas y los programas puede hacer que parezca que la metáfora del computador es neutral respecto de la TRM. Pero cuando uno piensa en la constitución de los procesos mentales, la conexión entre la idea de que son computacionales y la idea de que hay un lenguaje de pensamiento, se torna evidente de inmediato.[9]

8. Como Dennett, estoy suponiendo a los fines del argumento que la máquina *tiene* pensamientos y procesos mentales; nada depende de esto, ya que podríamos, por supuesto, haber tenido la misma discusión acerca de las personas.

9. Ahora podemos ver qué cosa decir respecto del viejo chiste filosófico acerca de la ley de Kepler. Lo que se alega es que la metodología intencionalista permite inferir de 'La conducta de *x* obedece a la regla *r*', '*r* es una regla que *x* se representa explícitamente'. Se supone que la dificultad consiste en que esto permite inferir de 'El movimiento de los planetas obedece a la ley de Kepler', alguna versión astronómica del lenguaje del pensamiento.

¿Qué ocurre con los ejemplos subpersonales en los que *tenemos* ejemplificaciones de representaciones mentales sin ejemplificaciones de actitudes? Las explicaciones de creencias/deseos de sentido común resultan reivindicadas si la psicología científica está comprometida ontológicamente con deseos y creencias. Pero *no* se requiere además que el inventario de actitudes proposicionales de la psicología de sentido común deba agotar una clase natural. Sería asombroso si lo hiciera. ¿Cómo podría el sentido común saber todo eso? Lo que es importante respecto de la TRM —lo que hace de la TRM una reivindicación de la psicología intuitiva de creencias/deseos— no es que escoja una clase que es coextensiva con las actitudes proposicionales. La TRM muestra cómo podrían tener poderes causales los estados intencionales; justamente, el aspecto del realismo intencional de sentido común que parecía más misterioso desde un punto de vista metafísico.

La física molecular reivindica la taxonomía intuitiva de líquidos y sólidos para los objetos de tamaño mediano. Pero la clase más cercana a los líquidos que la física molecular reconoce incluye algunos que el sentido común no reconocería; el vidrio, por ejemplo. ¿Y qué?

Tanto mejor para la TRM; tanto mejor, también, para este capítulo. Prima facie hay un argumento fuerte a favor de la explicación de creencias/deseos de sentido común. El sentido común sería reivindicado si alguna buena teoría de la mente probara estar comprometida con entidades que —como las actitudes— son semánticamente evaluables y están etiológicamente involucradas, a la vez. La TRM parece ser una teoría de la mente así comprometida; por lo tanto, si la TRM es verdadera, el sentido común es reivindicado. Por cierto que la TRM necesita producir un ejemplo empírico; necesitamos buenas explicaciones, confirmadas independientemente, de los procesos mentales en tanto que sucesiones causales de transformaciones de representaciones mentales. La psicología cognitiva moderna está dedicada, prácticamente en su totalidad, a idear y a confirmar tales explicaciones...

TRADUCTORAS: Ana C. Couló, María C. González y Nora Stigol.
REVISION TÉCNICA: Eduardo Rabossi.

Pero, de hecho, *no* se supone ningún principio de inferencia tal. Lo que garantiza la hipótesis de que *r* está explícitamente representada no es la mera conducta de acuerdo con *r*; es una etiología de acuerdo con la cual *r* figura como el contenido de uno de los estados intencionales cuyas ejemplificaciones son causalmente responsables de la conducta de *x*. Y, por supuesto, *no* es parte de la historia etiológica de los movimientos de los planetas que la ley de Kepler les acaece cuando ellos se mueven.

REFERENCIAS BIBLIOGRÁFICAS

Churchland, P. M.: (1982) (*EMPA*) "Eliminative Materialism and Propositional Attitudes", *Journal of Philosophy* 78, nº 2.

Cummins, R.: (1982) (*IMM*) "The Internal Manual Model of Psychological Explanation", *Cognition and Brain Theory* 5, nº 3.

Demoupoulos, W.y Matthews, R.: (1983) (*HGMR*) "On the Hypothesis That Grammars Are Mentally Represented". *Behavioral and Brain Sciences* 3.

Dennett, D.: (1981) (*CCC*) "A Cure for the Common Code?", en *Brainstorms*. Cambridge, Mass., MIT Press.

Field, H.: (1978) (*MR*) "Mental Representation", *Erkenntnis* 13, nº 1.

Fodor, J.: (1968) (*ATK*) "The Appeal to Tacit Knowledge in Psychological Explanation", *Journal of Philosophy* 65, nº 20 (incluido en Fodor, J., *Representations*).

Fodor, J.: (1974) (*SS*) "Special Sciences", *Synthèse* 28 (incluido en Fodor, J., *Representations*).

Fodor, J.: (1975) (*LOT*) *The Language of Thought*. Nueva York, Thomas Y. Crowell.

Fodor, J.: (1978) (*PA*) "Propositional Attitudes", *The Monist* 64, nº 4 (incluido en Fodor, J., *Representations*).

Fodor, J.: (1981) (*R*) *Representations*. Cambridge, Mass., MIT Press.

Fodor, J.: (*SSA*) "Something on the State of the Art", Introducción a Fodor, J., *Representations*.

Fodor, J.: (1983) (*MOM*) *The Modularity of Mind*. Cambridge, Mass., MIT Press.

Fodor, J.: (1983) (*P*) *Psychosemantics*. Cambridge, Mass., MIT Press.

Lycan, W.: (1986) (*TB*) "Tacit Belief", en Bogdan, R. (comp.), *Belief*. Oxford. Oxford University Press.

Matthews, R.: (1984) (*TWR*) "Troubles with Representationalism", *Social Research* 51, nº 4.

Stabler, E.: (1983) (*HAGR*) "How Are Grammars Represented?", *Behavioral and Brain Sciences* 3.

Stich, S.: (1983) (*FFPCS*) *From Folk Psychology to Cognitive Science*. Cambridge, Mass., MIT Press.

III

LA NATURALEZA Y LA VIABILIDAD DE LOS MODELOS FUNCIONALISTAS

CAPÍTULO 4

LAS DIFICULTADES DEL FUNCIONALISMO (SELECCIÓN) *

Ned Block

1.0 Funcionalismo, conductismo y fisicalismo

El punto de vista funcionalista acerca de la naturaleza de la mente es, en este momento, ampliamente aceptado.[1] Al igual que el conductismo y el fisicalismo, el funcionalismo pretende responder a la pregunta "¿Qué son los estados mentales?". Me ocuparé de las formulaciones del funcionalismo *via* la tesis de la identidad. Ellas dicen, por ejemplo, que el dolor es un estado funcional, así como las formulaciones del fisicalismo *via* la tesis de la identidad dicen que el dolor es un estado físico.

Comenzaré describiendo al funcionalismo y esbozando la crítica funcionalista al conductismo y al fisicalismo. Luego argumentaré que las dificultades atribuidas por el funcionalismo al conductismo y al fisicalismo infectan también al funcionalismo.

Una caracterización del funcionalismo que es probable que sea lo suficientemente vaga como para ser aceptada por la mayoría de los funcionalistas, es la siguiente: cada tipo [*type*] de estado mental es un estado que consiste en una disposición a actuar de ciertas maneras *y a tener ciertos estados mentales*, dados ciertos *inputs* sensoriales y ciertos estados mentales. Expuesto de este modo, el funcionalismo puede verse como una nueva encarnación del conductismo. El conductismo identifica a los estados mentales con disposiciones a actuar de ciertas maneras en ciertas situaciones de *input*. Pero como han señalado sus críticos

* "Troubles with Functionalism", en *Perception and Cognition. Minnesota Studies in the Philosophy of Science. Vol. IX*, compilado por W. Savage, 1978. Con autorización del autor y de Minnesota University Press.

1. Vease Fodor; 1965; Lewis, 1972; Putnam, 1966, 1967, 1970, 1975a; Armstrong, 1968; quizá Sellars, 1968; quizá Dennett, 1969, 1978b; Nelson, 1969, 1975 (pero véase también Nelson, 1976); Pitcher, 1971; Smart, 1971; Block y Fodor, 1972; Harman, 1973; Grice, 1975; Shoemaker, 1975; Wiggins, 1975.

(Chisholm, 1957; Geach, 1957; Putnam, 1963), desear G como meta, no puede identificarse con, digamos, la disposición a hacer A en circunstancias de *input* en las cuales A conduce a G, puesto que, después de todo, el agente podría no *saber* que A conduce a G y de este modo podría no estar dispuesto a hacer A. El funcionalismo reemplaza a los "*inputs* sensoriales" conductistas por "*inputs* sensoriales y estados mentales", y el funcionalismo reemplaza las "disposiciones a actuar" conductistas por "disposiciones a actuar y tener ciertos estados mentales". Los funcionalistas quieren individuar causalmente a los estados mentales, y puesto que los estados mentales tienen causas y efectos mentales tanto como causas sensoriales y efectos conductuales, los funcionalistas individuan a los estados mentales, en parte, en términos de las relaciones causales con otros estados mentales. Una consecuencia de esta diferencia entre el funcionalismo y el conductismo es que existen organismos posibles que de acuerdo con el conductismo tienen estados mentales pero que, de acuerdo con el funcionalismo, no los tienen.

De tal modo, las condiciones necesarias de lo mental que el funcionalismo postula son, en un aspecto, más fuertes que las postuladas por el conductismo. De acuerdo con el conductismo, es necesario y suficiente para desear que G, que un sistema sea caracterizado por un cierto conjunto (quizás infinito) de relaciones *input-output*; es decir, de acuerdo con el conductismo, un sistema desea que G en el caso de que un cierto conjunto de condicionales de la forma "Emitirá O dado I" sea verdadero de él. Sin embargo, de acuerdo con el funcionalismo, un sistema podría tener esas relaciones *input-output* aunque no deseara que G; porque de acuerdo con el funcionalismo, que un sistema desee que G depende de que tal sistema tenga estados internos que tienen ciertas relaciones causales con otros estados internos (y con *inputs* y *outputs*). Puesto que el conductismo no apela al requerimiento de "estado interno", existen sistemas posibles de los cuales el conductismo afirma y el funcionalismo niega que tengan estados mentales.[2] Una manera de enunciar esto es que, de acuerdo con el funcionalismo, el conductismo peca de *liberalismo*, al adscribir propiedades mentales a cosas que de hecho no las tienen.

A pesar de la diferencia entre funcionalismo y conductismo que acabamos de esbozar, no es necesario que los funcionalistas y los conductistas no compartan un mismo espíritu.[3] Shoemaker (1975), por ejemplo,

2. La inversa es también verdadera.

3. Ciertamente, si uno define 'conductismo' como el punto de vista de que los tér-

dice: "En una de sus interpretaciones, el funcionalismo en la filosofía de la mente es la doctrina de que los términos mentales o psicológicos son, en principio, eliminables de una cierta manera" (págs. 306-7). Los funcionalistas han tendido a tratar a los términos de estado-mental en una caracterización funcional de un estado mental, de manera muy diferente de la de los términos de *input* y de *output*. Así, en la versión más simple de la teoría, en términos de máquina de Turing (Putnam, 1967; Block y Fodor, 1972), los estados mentales se identifican con la totalidad de los estados de máquina-de-Turing, los que se definen a sí mismos *implícitamente* mediante una tabla de máquina que menciona *explícitamente* los *inputs* y los *outputs*, descriptos de manera no-mentalista.

Según la versión del funcionalismo que ofrece Lewis, los términos de estado-mental se definen por medio de una modificación del método de Ramsey, de manera tal que elimina el uso esencial de terminología mental de las definiciones pero no elimina la terminología de *input* y *output*. Es decir, 'dolor' se define como sinónimo de una descripción definida que contiene términos de *input* y de *output* pero no terminología mental (véase Lewis, 1972).

Además, el funcionalismo tanto en las versiones de máquina como en las que no son de máquina, insistió de modo típico en que las caracterizaciones de los estados mentales deberían contener descripciones de *inputs* y de *outputs* en lenguaje *físico*. Armstrong (1968), por ejemplo, dice:

> Podemos distinguir entre 'conducta física', que refiere a cualquier acción o pasión del cuerpo meramente física, y 'conducta propiamente dicha', que implica relación con la mente... Ahora bien, si en nuestra fórmula ["estado de la persona apto para producir cierta clase de conducta"] 'conducta' significara 'conducta propiamente dicha', entonces estaríamos dando una explicación de los conceptos mentales en términos de un concepto que ya presupone la mentalidad, lo cual sería circular. De este modo, resulta claro que en nuestra fórmula 'conducta' tiene que significar 'conducta física' (pág. 84).

En consecuencia, puede decirse que el funcionalismo "ubica" a los estados mentales sólo en la periferia, es decir, mediante la especificación física, o al menos no-mental, de *inputs* y de *outputs*. Una tesis básica de este artículo es que, a causa de este rasgo, el funcionalismo no puede

minos mentales pueden definirse en términos no-mentales, entonces el funcionalismo *es* una versión del conductismo.

evitar el tipo de problema por el cual condena correctamente al conductismo. El funcionalismo también peca de liberalismo, por razones muy similares a las del conductismo. Sin embargo, a diferencia del conductismo, el funcionalismo puede ser alterado con naturalidad para evitar el liberalismo; pero sólo al precio de fracasar de manera igualmente ignominiosa.

El fracaso del que hablo es el que el funcionalismo atribuye al *fisicalismo*. Por 'fisicalismo' significo la doctrina de que el dolor, por ejemplo, es idéntico a un estado físico (o fisiológico).[4] Como muchos filósofos argumentaron (notablemente Fodor, 1965; Putnam, 1966; véase también Block y Fodor, 1972), si el funcionalismo es verdadero, es probable que el fisicalismo sea falso. Este punto se ve más claramente en relación con las versiones del funcionalismo en términos de máquina de Turing. Cualquier máquina de Turing abstracta dada puede realizarse en una amplia variedad de dispositivos físicos; es plausible, por cierto, que dada una correspondencia putativa entre un estado de máquina de Turing y un estado de configuración física (o fisiológica), habrá una realización posible de la máquina de Turing que proporcionará un contraejemplo a esa correspondencia. (Véase Kalke, 1969; Gendron, 1971, y Mucciolo, 1974, para argumentos en contra no-convincentes; véase también Kim, 1972.) En consecuencia, si dolor es un estado funcional no puede, por ejemplo, ser un estado cerebral, porque las criaturas sin cerebro pueden realizar la misma máquina de Turing que las criaturas con cerebro.

Tengo que destacar que el argumento funcionalista contra el fisicalismo no apela, meramente, al hecho de que una máquina de Turing abstracta pueda ser realizada mediante sistemas de *composición material* diferente (madera, metal, vidrio, etcétera). Argumentar de este modo

4. Estado tipo [*state type*], no estado caso [*token*]. A lo largo de este artículo, entenderé por 'fisicalismo' la doctrina que dice que cada tipo distinto de estado mental es idéntico a un tipo distinto de estado físico; por ejemplo, dolor (el universal) es un estado físico. El fisicalismo de casos, por otra parte, es la doctrina (más débil) de que cada dolor particular fechable es un estado físico de uno u otro tipo. El funcionalismo muestra que el fisicalismo de tipos es falso, pero no muestra que el fisicalismo de casos sea falso. Por 'fisicalismo' entiendo fisicalismo de *primer orden*; la doctrina de que, por ejemplo, la propiedad de tener dolor es una propiedad física de primer orden (en el sentido Russell-Whitehead). (Una propiedad de primer orden es aquella cuya definición no requiere cuantificación sobre propiedades; una propiedad de segundo orden es aquella cuya definición requiere cuantificación sobre propiedades de primer orden, y sobre ninguna otra propiedad.) La afirmación de que tener un dolor es una propiedad física de segundo orden es en realidad una forma (fisicalista) de funcionalismo. Véase Putnam, 1970.

sería como argumentar que la temperatura no puede ser una magnitud microfísica porque la misma temperatura puede ser poseída por objetos con *diferentes* estructuras microfísicas (Kim, 1972). Los objetos con diferentes estructuras microfísicas tal como objetos hechos de madera, de metal, de vidrio, etcétera, pueden tener muchas propiedades microfísicas interesantes en común, tal como una energía cinética molecular del mismo valor promedio. Más bien, el argumento funcionalista contra el fisicalismo es que es difícil ver cómo *podría haber* una propiedad física de primer orden no-trivial (ver la nota 4) en común con todas las realizaciones físicas posibles de un estado de máquina de Turing dado y sólo con ellas. ¡Trátese de pensar en un candidato remotamente plausible! Al menos, la prueba de cómo concebir uno recae en quienes piensan que tales propiedades físicas son concebibles.

Una manera de expresar este punto es que de acuerdo con el funcionalismo, el fisicalismo es una teoría *chauvinista*: niega propiedades mentales a sistemas que de hecho las tienen. Al decir, por ejemplo, que los estados mentales son estados cerebrales los fisicalistas excluyen injustamente a las pobres criaturas carentes de cerebro que, sin embargo, tienen mente.

Un segundo punto importante de este trabajo es que el argumento mismo que el funcionalismo usa para condenar al fisicalismo puede aplicarse con igual éxito contra el funcionalismo; ciertamente cualquier versión del funcionalismo que evite el liberalismo cae, como el fisicalismo, en el chauvinismo.

Este artículo tiene tres partes. La primera argumenta que el funcionalismo es culpable de liberalismo; la segunda, que una manera de modificar al funcionalismo para evitar el liberalismo es unirlo más firmemente a la psicología empírica, y la tercera, que ninguna versión del funcionalismo puede evitar tanto el liberalismo como el chauvinismo.

1.1 Algo más acerca de lo que el funcionalismo es

Una manera de ordenar la desconcertante variedad de teorías funcionalistas consiste en distinguir entre aquellas que se exponen en términos de una máquina de Turing y aquellas que no.

Una tabla de máquina de Turing lista un conjunto finito de estados de tabla-de-máquina, $S_1 \ldots S_n$; de *inputs*, $I_1 \ldots I_m$; y de *outputs*, $O_1 \ldots O_p$. La tabla especifica un conjunto de condicionales de la forma: si la máquina está en el estado S_i y recibe el *input* I_j, emite el *output* O_k y pasa al estado S_l. Es decir, dado cualquier estado y cualquier *input*, la

tabla especifica un *output* y un estado siguiente. Cualquier sistema con un conjunto de *inputs*, *outputs* y estados relacionados de la manera especificada por la tabla, es descripto por la tabla y es una realización del autómata abstracto especificado por la tabla.

Para tener el poder de computar cualquier función recursiva, una máquina de Turing tiene que ser capaz de controlar su *input* de ciertas maneras. En las formulaciones estándar se considera que el *output* de una máquina de Turing tiene dos componentes. Imprime un símbolo sobre una cinta, luego corre la cinta y pone, así, un nuevo símbolo ante la vista del lector del *input*. Para que una máquina de Turing tenga poder completo, la cinta tiene que ser infinita, en al menos una dirección y corrible en ambas direcciones. Si la máquina no tiene control sobre la cinta, es un "transductor finito" ["*finite transducer*"], una máquina de Turing muy limitada. No es necesario considerar que los transductores finitos tengan una cinta. Quienes creen que el funcionalismo de máquina es verdadero tienen que suponer que la cuestión acerca del poder que tenemos como autómatas, es una cuestión empírica sustantiva. Si somos máquinas de Turing de "poder completo", el entorno [*environment*] tiene que constituir parte de la cinta...

Una versión muy simple del funcionalismo de máquina (Block y Fodor, 1972) sostiene que cada sistema que tiene estados mentales es descripto por al menos una tabla de máquina de Turing especificable, y que cada tipo de estado mental del sistema es idéntico a uno de los estados de la tabla-de-máquina [*machine table*]. Consideremos, por ejemplo, la máquina de Turing descripta en el siguiente cuadro (cf. Nelson, 1975):

	S_1	S_2
moneda de $ 0,05 [*nickel*] *input*	No emite *output* Pasa a S_2	Emite una gaseosa Pasa a S_1
moneda de $ 0,10 [*dime*] *input*	Emite una gaseosa Permanece en S_1	Emite una gaseosa y una moneda de $ 0.05 Pasa a S_1

Se puede contar con una descripción cruda de la versión simple del

funcionalismo de máquina si se considera la afirmación de que S_1 = deseo-moneda de 0,10 [*nickel-desire*], y S_2 = deseo-moneda de 0,05 [*dime-desire*]. Por supuesto que ningún funcionalista sostendría que una máquina de ese tipo desee algo. Más bien, la versión simple de funcionalismo de máquina descripta arriba formula un planteo análogo respecto de una hipotética tabla de máquina mucho más compleja. Adviértase que el funcionalismo de máquina especifica explícitamente a los *inputs* y *outputs* e implícitamente a los estados internos (Putnam [1967, pág. 434] afirma: "Para decirlo una vez más, los S_i se especifican sólo *implícitamente* por la descripción, es decir, se especifican *sólo* por el conjunto de probabilidades de transición dado en la tabla de máquina"). Un dispositivo tiene que aceptar monedas de $ 0,05 y de $ 0,10 como *inputs* y devolver monedas de $ 0,05 y gaseosas como *outputs*, para ser descripto por esa tabla de máquina. Pero los estados S_1 y S_2 pueden ser virtualmente de cualquier naturaleza (aun de naturaleza no-física), en tanto que esa naturaleza conecte a los estados entre sí y con los *inputs* y *outputs* especificados en la tabla de máquina. Todo lo que se nos dice de S_1 y S_2 son esas relaciones; así, puede decirse que el funcionalismo de máquina reduce la mentalidad a estructuras *input-output*. Este ejemplo debería sugerir la fuerza del argumento funcionalista contra el fisicalismo. ¡Trátese de pensar en una propiedad física de primer orden (véase la nota 4) que pueda ser compartida por todas las realizaciones de esta tabla de máquina y sólo por ella!

También se puede caracterizar a los funcionalistas según que consideren a las identidades funcionales como parte de una psicología a priori o de una psicología empírica... Los funcionalistas a priori (por ejemplo, Smart, Armstrong, Lewis, Shoemaker), son los herederos de los conductistas lógicos. Tienden a considerar a los análisis funcionales como análisis de los significados de los términos mentales, mientras que los funcionalistas empíricos (por ejemplo, Fodor, Putnam, Harman) consideran a los análisis funcionales como hipótesis científicas substantivas. En lo que sigue, referiré al primer punto de vista como 'Funcionalismo' y al último como 'Psicofuncionalismo'. (Usaré 'funcionalismo', con 'f' minúscula, como neutral entre Funcionalismo y Psicofuncionalismo. Cuando distinga entre Funcionalismo y Psicofuncionalismo usaré siempre letras mayúsculas.)

El Funcionalismo y el Psicofuncionalismo y la diferencia entre ellos pueden clarificarse en términos de la noción de la oración Ramsey [*Ramsey sentence*] de una teoría psicológica. Los términos de estado-mental [*mental-state terms*] que aparecen en una teoría psicoló-

gica pueden definirse de varias maneras mediante la oración Ramsey de la teoría... Todas las teorías de la identidad de estado funcional [*functional state identity theories*] pueden entenderse como definiendo un conjunto de estados funcionales ... mediante la oración Ramsey de una teoría psicológica, correspondiendo un estado funcional a cada estado mental. El estado funcional que corresponde a dolor se llamará 'el correlato funcional Ramsey' [*Ramsey functional correlate*] de dolor, con respecto a la teoría psicológica. En términos de la noción de correlato funcional Ramsey de una teoría, la distinción entre Funcionalismo y Psicofuncionalismo puede definirse como sigue: el Funcionalismo identifica el estado mental S con el correlato funcional Ramsey de S, con respecto a una teoría psicológica de *sentido común*; el Psicofuncionalismo identifica S con el correlato funcional Ramsey de S con respecto a una teoría psicológica *científica*.

Esta diferencia entre Funcionalismo y Psicofuncionalismo da origen a una diferencia en la especificación de los *inputs* y *outputs*. Los Funcionalistas están limitados a especificar *inputs* y *outputs* que sean una parte plausible del conocimiento de sentido común [*common-sense knowledge*]; los Psicofuncionalistas no tienen tal limitación. Si bien ambos grupos insisten en la especificación física —o al menos no-mental— de *inputs* y de *outputs*, los Funcionalistas precisan clasificaciones externamente observables (tales como *inputs* caracterizados en términos de objetos presentes en la vecindad del organismo, *outputs* en términos de movimientos de partes del cuerpo). Los Psicofuncionalistas, en cambio, tienen la opción de especificar *inputs* y *outputs* en términos de parámetros internos tales como señales en las neuronas de *input* y de *output*...

Sea T una teoría psicológica de sentido común o bien de psicología científica [*psychological theory of either common-sense or scientific knowledge*]. T puede contener generalizaciones de la forma: quien quiera que esté en el estado w y reciba el *input* x emite el *output* y, y pasa al estado z. Escribamos T como

$$T\,(S_1 \ldots S_n, I_1 \ldots I_k, O_1 \ldots O_m)$$

en donde las S son estados mentales, las I son *inputs* y las O son *outputs*. Las 'S' han de entenderse como *constantes* de estado mental, tal como 'dolor', no como variables; lo mismo vale para las 'I' y las 'O'. Así, uno podría también escribir T como

T (dolor..., luz de 400 nanómetros entrando por el ojo izquierdo..., dedo gordo del pie izquierdo se mueve 1 centímetro á la izquierda...)

Para obtener la oración Ramsey de T, reemplácense por variables los términos correspondientes a estados mentales —pero *no los términos correspondientes a* inputs *y* outputs—, y prefíjese un cuantificador existencial para cada variable.

$$\exists F_1 \ldots \exists F_n T(F_1 \ldots F_n, I_1 \ldots I_k, O_1 \ldots O_m)$$

Si 'F_{17}' es la variable que reemplazó a la palabra 'dolor' cuando se formó la oración Ramsey, entonces podemos definir al dolor, en términos de la oración Ramsey, como sigue:

$$x \text{ tiene (siente) dolor } [x \textit{ is in pain}] \leftrightarrow \exists F_1 \ldots \exists F_n \; T[(F_1 \ldots F_n, I_1 \ldots I_k, O_1 \ldots O_m) \;\&\; x \text{ tiene } F_{17}]$$

El correlato funcional Ramsey de dolor es la propiedad expresada por el predicado del lado derecho de este bicondicional. Nótese que este predicado contiene constantes de *input* y de *output*, pero no constantes [de términos] mentales puesto que las constantes [de términos] mentales fueron reemplazadas por variables. El correlato funcional Ramsey de dolor es definido en términos de *inputs* y *outputs*, pero no en términos mentales.

Por ejemplo, sea T la teoría acerca de que el dolor es causado por daño en la piel y es causa de preocupación y de la proferencia de "*ouch*", y que la preocupación causa a su vez fruncir el entrecejo. Entonces la definición Ramsey sería:

> x tiene (siente) dolor ↔ Hay 2 estados (propiedades), el primero de los cuales es causado por daño en la piel y causa la proferencia de "*ouch*" y del segundo estado, y el segundo estado causa fruncir el entrecejo, y x está en el primer estado.

El correlato funcional Ramsey de dolor con respecto a esta "teoría" es la propiedad de estar en un estado que es causado por daño en la piel y que causa la proferencia de "*ouch*" y otro estado que causa a su vez fruncir el entrecejo. (Nótese que las palabras 'dolor' y 'preocupación' han sido reemplazadas por variables, pero los términos de *input* y de *output* no.)

El correlato funcional Ramsey de un estado S es un estado que tiene mucho en común con S. Específicamente, S y su correlato funcional Ramsey comparten las propiedades *estructurales* especificadas por la teoría T. Pero, existen dos razones por las cuales es natural suponer que S y su correlato funcional Ramsey serán distintos. Primero, el correlato funcional Ramsey de S con respecto a T puede "incluir", a lo sumo, aquellos aspectos de S que están relevados [*captured*] por T; los aspectos que no estén relevados por T, quedan afuera. Segundo, el correlato funcional Ramsey podría dejar también a un lado algo de lo que T releva porque la definición Ramsey no contiene el vocabulario "teórico" de T. La teoría tomada como ejemplo en el último parágrafo es verdadera sólo de los organismos que sienten-dolor [*pain-feeling organisms*]; y lo es trivialmente, en virtud de su uso de la palabra 'dolor'. Sin embargo, el predicado que expresa el correlato funcional Ramsey no contiene esa palabra (puesto que fue reemplazada por una variable), y así puede ser verdadera de cosas que no sienten dolor. Sería sencillo construir una máquina simple que tenga piel artificial, una ceja, una cinta con la grabación de "*ouch*" y dos estados que satisfagan las relaciones causales mencionadas, pero que no sienta dolor.

La hipótesis fuerte del funcionalismo es que para *alguna* teoría psicológica, la suposición natural de que un estado y su correlato funcional Ramsey sean distintos, es falsa. El Funcionalismo dice que existe una teoría tal que dolor, por ejemplo, *es* el correlato funcional Ramsey respecto de ella.

Un último punto preliminar: he dado la impresión equivocada de que el funcionalismo identifica *todos* los estados mentales con estados funcionales. Una versión tal del funcionalismo es obviamente demasiado fuerte. Sea X una réplica célula-por-célula de uno, recién creada (la cual, por supuesto, es funcionalmente equivalente a uno). Quizás uno recuerde haber celebrado su *bar-mitzva*. Pero X no recuerda haber celebrado su *bar-mitzva*, puesto que X nunca lo tuvo. Por cierto, algo puede ser funcionalmente equivalente a uno pero no saber lo que uno sabe, o [verbo], lo que uno [verbo], para una amplia variedad de verbos de "éxito" ["*success*"]. Peor aún, si Putnam (1975b) está en lo correcto al decir que "los significados no están en la cabeza", sistemas funcionalmente equivalentes a uno pueden no tener, por razones similares, muchas de las demás actitudes proposicionales de uno. Supongamos que uno cree que el agua es húmeda. De acuerdo con ciertos argumentos plausibles presentados por Putnam y Kripke, una condición para la posibilidad de que uno crea que el agua es húmeda es un cierto tipo de cone-

xión causal entre uno y el agua. Nuestro "gemelo" en la Tierra Gemela que está conectado de una manera similar a XYZ pero no a H_2O, no creería que el agua es húmeda.

Si el funcionalismo ha de ser defendido, tiene que ser interpretado como aplicándose solamente a una subclase de estados mentales: aquellos estados mentales "estrechos" ["*narrow*"] tal que las condiciones de verdad para su aplicación estén en algún sentido "dentro de la persona". Pero aun suponiendo que una noción del carácter estrecho de los estados psicológicos pueda ser formulada satisfactoriamente, el interés del funcionalismo puede disminuir a causa de esa limitación. Menciono este problema sólo para dejarlo a un lado.

Consideraré al funcionalismo como una doctrina acerca de estados mentales "estrechos".

1.2 Robots de cabeza homuncular [*homunculi-headed robots*]

En esta sección describiré una clase de estratagemas [*devices*] que, prima facie, ponen en un aprieto a todas las versiones del funcionalismo, dado que indican que el funcionalismo peca de liberalismo al clasificar sistemas que carecen de mentalidad, como teniendo mentalidad.

Consideremos la versión simple del funcionalismo de máquina ya descripto. Dice que cada sistema que tiene estados mentales es descripto por al menos una tabla de máquina de Turing de un cierto tipo y que cada estado mental del sistema es idéntico a uno de los estados de tabla-de-máquina especificados por la tabla de máquina. Consideraré que los *inputs* y los *outputs* son especificados mediante descripciones de impulsos neurales en los órganos sensoriales y mediante neuronas de *output* motor. No debe considerarse que lo que se va a decir vale para el Psicofuncionalismo y no para el Funcionalismo. Tal como señalé, toda versión del funcionalismo supone *alguna* especificación de *inputs* y de *outputs*. Una especificación Funcionalista serviría lo mismo para nuestros fines.

Imaginemos un cuerpo externamente similar a un cuerpo humano, digamos como el de uno, pero internamente muy diferente. Las neuronas asociadas a los órganos sensoriales se conectan con una hilera de luces ubicada en una cavidad vacía de la cabeza. Un conjunto de botones está conectado con las neuronas de *output* motoras. Dentro de la cavidad reside un grupo de hombrecitos. Cada uno tiene una tarea muy sencilla: implementa una "casilla" ["*square*"] de una tabla de máquina adecuada que lo describe a uno. Sobre una pared hay una cartelera en

la que está colocada una tarjeta de estado [*state card*]; es decir, una tarjeta que tiene un símbolo que designa a uno de los estados especificados en la tabla de máquina. He aquí lo que los hombrecitos hacen. Supongamos que la tarjeta tiene una 'G'. Esto alerta al hombrecito que implementa los casilleros G. Los hombrecitos se autodenominan 'hombres-G'. Supongamos que la luz que representa al *input* I_{17} está encendida. Uno de los hombres-G sólo tiene la siguiente tarea: cuando la tarjeta dice 'G' y la luz I_{17} está encendida, él presiona el botón de *output* O_{191} y cambia la tarjeta de estado a 'M'. Este hombre-G es llamado a ejercitar su tarea sólo en raras ocasiones. A pesar del bajo nivel de inteligencia que se requiere de cada hombrecito, el sistema, como un todo, se las arregla para simularlo a uno, porque la organización funcional para cuya realización se los entrenó, es la de uno. Una máquina de Turing puede ser representada como un conjunto finito de cuádruplas (o quíntuplas, si el *output* es dividido en dos partes): estado actual, *input* actual, estado próximo y *output* próximo. Cada hombrecito tiene una tarea que corresponde a una única cuádrupla. A través de los esfuerzos de los hombrecitos el sistema realiza la misma (razonablemente adecuada) tabla de máquina que uno y, de tal modo, es funcionalmente equivalente a uno.[5]

Describiré una versión de la simulación de cabeza homuncular, que tiene más probabilidades de ser nomológicamente posible. ¿Cuántos homúnculos se requieren? Quizás un par de miles de millones sea suficiente.

Supongamos que convertimos al gobierno de China al funcionalismo y convencemos a sus funcionarios... para que realicen una mente humana durante una hora. Proporcionamos a cada una de los miles de millones de personas de China (elijo a China porque tiene un par de miles de millones de habitantes) un radiotransmisor de doble canal especialmente diseñado, que las conecta de manera apropiada con otras personas y con el cuerpo artificial mencionado en el ejemplo anterior. Reemplazamos a cada uno de los hombrecitos por un ciudadano chino y su radio-transmisor. En vez de una cartelera tenemos letras desplegadas en una serie de satélites ubicados de modo tal que puedan ser vistos desde cualquier lugar de China.

5. La idea básica de este ejemplo proviene de Putnam (1967). Estoy en deuda con Hartry Field por las conversaciones mantenidas sobre este tema. El intento de Putnam de evitar al funcionalismo el problema planteado por tales ejemplos es discutido en la sección 1.3 de este trabajo.

El sistema de un par de miles de millones de personas que se comunican entre sí más los satélites, desempeña el rol de un "cerebro" externo conectado a un cuerpo artificial mediante radiotransmisión. No hay nada absurdo acerca de una persona conectada a su cerebro mediante radiotransmisión. Llegará el día, quizás, en que nuestros cerebros sean periódicamente retirados para limpieza y reparación. Imaginemos que esto se hace primero tratando a las neuronas que acoplan al cerebro con el cuerpo con una sustancia química que les permita estirarse como bandas de goma, asegurando con ello que ninguna de las conexiones cuerpo-cerebro sea interrumpida. Muy pronto, hombres de negocio inteligentes descubren que pueden atraer más clientes reemplazando las neuronas estiradas por nexos de radiotransmisión, de manera que los cerebros puedan ser limpiados sin incomodar al cliente al tener que inmovilizar su cuerpo.

No es nada obvio que el sistema corporal chino sea físicamente imposible. Podría ser funcionalmente equivalente a uno por un tiempo breve, digamos una hora.

"Pero —alguien puede objetar— ¿cómo podría algo ser funcionalmente equivalente a mí por *una hora*? ¿No determina mi organización funcional, digamos, cómo reaccionaría si durante una semana lo único que hiciera fuera leer el *Reader's Digest*?" Recordemos que una tabla de máquina especifica un conjunto de condicionales de la forma: si la máquina está en S_i y recibe el *input* I_j, emite el *output* O_k y pasa a S_l. Estos condicionales tienen que entenderse *subjuntivamente*. Lo que le da a un sistema una organización funcional en un momento dado no es lo que *hace* en ese momento sino también los contrafácticos que son verdaderos de él en ese momento: lo que *hubiera* hecho (y lo que hubieran sido sus transiciones de estado) de haber tenido un *input* diferente o de haber estado en un estado diferente. Si es verdad de un sistema, en el tiempo t, que *obedecería* a una tabla de máquina dada sin importar en cuál de los estados esté y sin importar cuál de los *inputs* reciba, entonces el sistema es descripto, en t, mediante la tabla de máquina (y realiza en t al autómata abstracto especificado por la tabla), aun si existiera sólo por un instante. Durante la hora en la cual el sistema chino está "en funcionamiento" *tiene* un conjunto de *inputs*, *outputs* y estados de los cuales tales condicionales subjuntivos son verdaderos. Esto es lo que hace que cualquier computador realice el autómata abstracto que realiza.

Hay señales, por supuesto, a las que el sistema respondería y a las que uno no respondería, por ejemplo, a una interferencia masiva en la

radiotransmisión o a una inundación del río Yangtze. Tales eventos podrían causar un mal funcionamiento frustrando la simulación, como una bomba [*bomb*] en un computador puede hacer que el computador no realice la tabla de máquina para cuya realización fue construido. Pero así como el computador *sin* la bomba *puede* realizar la tabla de máquina, el sistema que se compone de personas y cuerpo artificial puede realizar la tabla de máquina en tanto no haya catástrofes que interfieran, tales como inundaciones, etcétera.

"Pero —alguien puede objetar— existe una diferencia entre una bomba en un computador y una bomba en el sistema chino, porque en el caso de este último (a diferencia del primero), los *inputs* especificados en la tabla de máquina pueden ser la causa del mal funcionamiento. La actividad neural inusual en los órganos sensoriales de los residentes de la provincia de Chungking ocasionada por una bomba o por una inundación del Yangtze, puede causar que el sistema se desordene."

Respuesta: la persona que dice a qué sistema se refiere tiene que decir qué señales valen como *inputs* y *outputs*. Yo tomo como *inputs* y como *outputs* sólo a la actividad neural en el cuerpo artificial conectado mediante radiotransmisión con los habitantes de China. Las señales neurales de los habitantes de Chungking cuentan tan poco como *input* de ese sistema, como la cinta de *input* atascada por un saboteador entre los contactos de relé en las entrañas de una computadora, cuenta como *input* de esa computadora.

Por supuesto, el objeto que se compone de los habitantes de China + el cuerpo artificial, tiene *otras* descripciones de máquina de Turing bajo las cuales las señales neurales en los habitantes de Chungking *contarían* como *inputs*. Ese nuevo sistema (esto es, el objeto de esa nueva descripción de máquina de Turing) no sería funcionalmente equivalente a uno. De modo similar, cualquier computador comercial puede ser redescripto de manera que permita que la cinta atascada en su interior cuente como *input*. Al describir un objeto como una máquina de Turing, uno traza una línea entre dentro y fuera. (Si sólo consideramos a los impulsos neurales como *inputs* y *outputs*, trazamos esa línea dentro del cuerpo; si sólo consideramos a las estimulaciones periféricas como *inputs*, ...trazamos esa línea en la piel.) Al describir al sistema chino como una máquina de Turing, he trazado la línea de tal manera que satisface un cierto tipo de descripción funcional, una [descripción] que *también* uno satisface y que, de acuerdo con el funcionalismo, justifica adscripciones de mentalidad. El Funcionalismo no sostiene que todo sistema mental tenga una tabla de máquina de un tipo tal que justifique

adscripciones de mentalidad con respecto a *toda* especificación de *inputs* y de *outputs*, sino más bien, sólo con respecto a *alguna* especificación

Objeción: el sistema chino trabajaría demasiado lentamente. El tipo de eventos y procesos con los que tenemos contacto normalmente, ocurrirían demasiado rápido para que el sistema los detectase. Así, no estaríamos en condiciones de conversar con él, jugar bridge con él, etcétera

Respuesta: resulta difícil ver por qué la escala temporal del sistema debe importar... ¿Es realmente contradictorio o sin sentido suponer que podríamos encontrar una raza de seres inteligentes con los cuales podríamos comunicarnos sólo a través de dispositivos tales como una cámara lenta [*time-lapse photography*]? Cuando observamos a esas criaturas, parecen casi inanimadas. Pero cuando vemos las películas en cámara lenta, las vemos conversando entre sí. Por cierto, encontramos que dicen que la única manera en que ellas pueden entendernos es viendo las películas en cámara lenta. Considerar a la escala temporal como lo más importante parece crudamente conductista...

Lo que hace del sistema con cabeza-homuncular recién descripto (considérese a los dos sistemas como variantes de un único sistema) un contraejemplo posible del funcionalismo (de máquina), es que existe la duda, prima facie, de que tenga estados mentales, especialmente de que tenga lo que los filósofos han llamado, de diversos modos, "estados cualitativos", "vivencias puras" ["*raw feels*"] o "cualidades fenomenológicas inmediatas". (Alguien pregunta: ¿qué es lo que los filósofos han llamado estados cualitativos? Yo respondo bromeando sólo a medias: como dijo Louis Armstrong cuando le preguntaron qué es el *jazz*, "Si usted me lo tiene que preguntar, nunca podrá llegar a saberlo".) En términos de Nagel (1974), existe la duda, prima facie, de que haya algo que sea cómo ser el sistema con cabeza-homuncular.[6]

1.3 La propuesta de Putnam

Una manera en que los funcionalistas pueden tratar de encarar el problema planteado por los contraejemplos que recurren a cabezas-homunculares, es apelando al recurso ad hoc de no darles cabida. Por ejemplo, un funcionalista podría estipular que dos sistemas no

6. Shoemaker (1975) argumenta (en respuesta a Block y Fodor, 1972) que los *qualia* ausentes son lógicamente imposibles; esto es, que es lógicamente imposible que dos sistemas estén en el mismo estado funcional y que, sin embargo, uno tenga un contenido cualitativo y el otro carezca de él.

pueden ser funcionalmente equivalentes si uno contiene partes con organizaciones funcionales características de los seres sintientes [*sentient beings*] y el otro no. En el artículo en que hipotetiza que el dolor es un estado funcional, Putnam estipula que "ningún organismo capaz de sentir dolor es susceptible de ser descompuesto en partes que separadamente posean Descripciones" (como el tipo de máquina de Turing que puede estar en el estado funcional que Putnam identifica con dolor). El propósito de esta condición es "excluir 'organismos' (si es que valen como tales) como los enjambres de abejas en tanto que experimentadores singulares de dolor" (Putnam, 1967, págs. 434-5).

Una manera de satisfacer el requisito de Putnam sería ésta: un organismo que es capaz de sentir-dolor no es susceptible de ser descompuesto en partes, *todas* las cuales tengan una organización funcional característica de los seres-sintientes. Pero esto no excluye mi ejemplo que apela a la cabeza-homuncular, dado que tiene partes no sintientes, tales como el cuerpo mecánico y los órganos sensoriales. No servirá irse al extremo opuesto y requerir que *ninguna* parte propia sea sintiente. De otro modo, las mujeres embarazadas y las personas con parásitos sintientes [*sentient parasites*] no podrían contar como organismos capaces de sentir dolor. Lo que parece ser importante para ejemplos como la simulación de cabeza-homuncular que he descripto, es que los seres sintientes *desempeñan un rol crucial* en dar a las cosas su organización funcional. Esto sugiere una versión de la propuesta de Putnam que requiere que un organismo capaz de sentir dolor tenga una cierta organización funcional y no tenga partes que (1) posean ellas mismas ese tipo de organización funcional y además (2) desempeñen un rol crucial en dar al sistema total su organización funcional.

Aunque esta propuesta involucra la noción vaga de "rol crucial", es lo suficientemente precisa para hacernos ver que no funcionará. Supongamos que existe una parte del universo que contiene una materia completamente diferente de la nuestra, una materia que es infinitamente divisible. En esa parte del universo hay criaturas inteligentes de muchos tamaños, incluso criaturas semejantes a los humanos pero mucho más pequeñas que nuestras partículas elementales. En una expedición intergaláctica esa gente descubre la existencia de nuestro tipo de materia. Por razones que ellos sólo conocen deciden dedicar los próximos cientos de años a producir, partiendo de *su* materia, sustancias con las características químicas y físicas (excepto en el nivel de partículas subelementales) de *nuestros* elementos. Construyen hordas de naves espaciales de diferentes variedades remedando el tamaño aproximado de nuestros

electrones, protones y otras partículas elementales, y pilotean las naves de manera de imitar el comportamiento de esas partículas elementales. Además, las naves contienen generadores para producir el tipo de radiación que producen las partículas elementales. Cada nave tiene un equipo de expertos en la naturaleza de nuestras partículas elementales. Hacen esto para producir inmensas (de acuerdo con nuestros estándares) masas de sustancias con las características químicas y físicas del oxígeno, el carbono, etcétera. Poco tiempo después de que han logrado su objetivo, uno sale de expedición a esa parte del universo y descubre el "oxígeno", el "carbono", etcétera. Ignorante de su verdadera naturaleza, uno establece una colonia, y usa esos "elementos" para cultivar plantas alimenticias, proporcionar "aire" para respirar, etcétera. Dado que las moléculas de uno son intercambiadas constantemente con el entorno, uno y los demás colonizadores (en un período de pocos años) llegamos a estar compuestos principalmente de la "materia" hecha de esa gente diminuta en sus naves espaciales. ¿Sería uno menos capaz de sentir dolor, de pensar, etcétera, sólo, porque la materia de la que está compuesto (y de la que dependen sus características) contiene seres que, en sí mismos, tienen una organización funcional típica de criaturas sintientes? Creo que no. Los mecanismos electroquímicos básicos mediante los cuales se lleva a cabo la sinapsis son ahora bastante bien comprendidos. Como se sabe, los cambios que no afectan a esos mecanismos electroquímicos no afectan al funcionamiento del cerebro y no afectan a la mentalidad. Los mecanismos electroquímicos en nuestras sinapsis no serían afectados por el cambio en nuestra materia.[7]

Resulta interesante comparar el ejemplo de la gente-hecha-de-partícula-elemental con los ejemplos del comienzo de capítulo que apelan a la cabeza-homuncular. Una conjetura natural acerca de la fuente de nuestra intuición de que las simulaciones descriptas inicialmente que apelan a la cabeza-homuncular carecen de mentalidad, es que tienen *demasiada* estructura mental interna. Los hombrecitos podrían a veces aburrirse, a veces excitarse. Podemos imaginar aun que deliberan acerca de la mejor manera de realizar la organización funcional dada y que hacen cambios con la intención de gozar de más tiempo libre. Pero el

7. Dado que hay una diferencia entre el rol de los hombrecitos al producir su organización funcional en la situación descripta y el rol de los homúnculos en las simulaciones que apelan a la-cabeza-homuncular, con que se inicia este trabajo, cabe presumir que la condición de Putnam podría ser reformulada de modo de excluir a los segundos sin excluir a los primeros. Pero esto sería una maniobra muy ad hoc.

ejemplo de la gente-hecha-de-partícula-elemental recién descripto, sugiere que esta primera conjetura es errónea. Lo que parece importante es *cómo* la mentalidad de las partes contribuye al funcionamiento del todo.

Hay una diferencia muy notable entre el ejemplo de la gente-hecha-de-partícula-elemental y los anteriores ejemplos de homúnculos. En el primero, el cambio que se produce en uno a medida que nos vamos infectando de homúnculos no es un cambio que produzca ninguna diferencia en nuestro procesamiento psicológico (es decir, procesamiento de información) o en nuestro procesamiento neurológico, sino sólo en nuestra microfísica. Ninguna de las técnicas propias de la psicología o de la neurofisiología humanas revelaría diferencia alguna en uno. Sin embargo, las simulaciones que apelan a la cabeza-homuncular, descriptas al comienzo del trabajo, no son cosas a las que se apliquen las teorías neurofisiológicas que son verdaderas de nosotros, y *si son interpretadas como* simulaciones *Funcionales* (más que como Psicofuncionales) no necesitan ser cosas a las que se apliquen las teorías psicológicas (procesamiento de información) verdaderas de nosotros. Esta diferencia sugiere que nuestras intuiciones están, en parte, controladas por el punto de vista, no del todo razonable, de que nuestros estados mentales dependen de que tengamos la psicología y/o la neurofisiología que tenemos. Así, algo que difiera marcadamente de nosotros en ambos aspectos (recuérdese que se trata de una simulación Funcional más que Psicofuncional) no debe suponerse que tenga mentalidad, sólo sobre la base de que ha sido diseñado para ser Funcionalmente equivalente a nosotros.

1.4 ¿Es la duda prima facie meramente prima facie?

El Argumento de los *Qualia* Ausentes [*Absent Qualia Argument*] descansó en una apelación a la intuición de que las simulaciones que apelan a la cabeza-homuncular carecían de mentalidad, o al menos, de *qualia*. He dicho que esta intuición dio origen a la duda, prima facie, de que el funcionalismo sea verdadero. Pero las intuiciones que no se apoyan en argumentos fundados [*principled*] difícilmente han de ser consideradas sólidas. Ciertamente, las intuiciones incompatibles con una teoría bien fundada, tal como la intuición precopernicana de que la Tierra no se mueve, felizmente desaparecen pronto. Aun en ámbitos como el de la lingüística, cuyos datos consisten principalmente de intuiciones,

a menudo se rechazan intuiciones tales como que las siguientes oraciones son no-gramaticales (sobre bases teóricas):

El caballo corrido pasó el establo cayó.

El muchacho la chica el gato mordió rasguñó murió.

Estas oraciones son, de hecho, gramaticales, aunque difíciles de procesar.[8]

Apelar a las intuiciones cuando se juzga la posesión de mentalidad es, sin embargo, *especialmente* sospechoso. *Ningún* mecanismo físico parece intuitivamente plausible como asiento de los *qualia*, y mucho menos un *cerebro*. ¿Es una guedeja de tembloroso material [*stuff*] gris más intuitivamente apropiada como asiento de los *qualia* que un grupo de hombrecitos? Si no lo es, quizás haya también una duda prima facie, acerca de los *qualia* de los sistemas con cabeza-cerebral [*brain-headed*].

Sin embargo, existe una diferencia muy importante entre los sistemas con cabeza-cerebral y los sistemas con cabeza-homuncular. Dado que sabemos que *nosotros somos sistemas con cabeza-cerebral* y que tenemos *qualia*, sabemos que los sistemas con cabeza-cerebral pueden tener *qualia*. Así, aunque carecemos de una teoría de los *qualia* que explique cómo es ello *posible*, tenemos una razón contundente para desechar toda duda prima facie que haya acerca de los *qualia* de los sistemas con cabeza-cerebral. Por supuesto que esto hace a mi argumento parcialmente *empírico*: depende del conocimiento que nos marca [*makes us tick*]. Pero dado que este es un conocimiento que de hecho poseemos, depender de tal conocimiento no debería ser considerado un defecto.[9]

Existe otra diferencia entre nuestras cabezas-de-carne-y-hueso y las cabezas-homunculares: éstos son sistemas diseñados para imitarnos,

8. Compárese la primera oración con 'El pescado comido en Boston apesta'. La razón de que sea difícil de procesar es que 'corrido' se lee de manera natural como activo más que como pasivo. Véase Fodor *et al.*, 1974, pág. 360. Para una discusión de por qué la segunda oración es gramatical, véase Fodor y Garrett, 1967; Bever, 1970, y Fodor *et al.*, 1974.

9. A menudo no podemos concebir cómo algo es posible porque carecemos de los conceptos teóricos relevantes. Por ejemplo, antes del descubrimiento de los mecanismos de duplicación genética, Haldane argumentó persuasivamente que ningún mecanismo físico concebible podría hacer ese trabajo. Estuvo en lo correcto. Pero en vez de argumentar que los científicos deberían desarrollar ideas que nos permitiesen concebir un mecanismo físico tal, concluyó que un mecanismo *no*-físico estaba involucrado. (Debo este ejemplo a Richard Boyd.)

pero nosotros no estamos diseñados para imitar nada (aquí me apoyo en otro hecho empírico). Este hecho cancela cualquier intento de argumentar sobre la base de una inferencia a la mejor explicación a favor de los *qualia* de cabezas-homunculares. La mejor explicación de los gritos y muecas de las cabezas-homunculares no son sus dolores, sino que fueron diseñadas para imitar nuestros gritos y muecas.

Algunas personas parecen sentir que la conducta compleja y sutil de las cabezas-homunculares (conducta tan compleja y sutil, aun tan "sensitiva" a los rasgos del entorno, humano y no-humano, como nuestra conducta) es por sí misma una razón suficiente para desechar la duda, prima facie, de que las cabezas-homunculares tengan *qualia*. Pero esto es crudo conductismo...

Mi argumento contra el Funcionalismo depende del siguiente principio: si una doctrina tiene una conclusión absurda para creer en la cual no hay una razón independiente, y si no hay manera de salvar el absurdo o de mostrar que es engañoso o irrelevante, y si no hay una buena razón para creer en la doctrina que lleva directamente al absurdo, entonces no se acepte la doctrina. Sostengo que no hay una razón independiente para creer en la mentalidad de una cabeza-homuncular, y sé que no hay manera de salvar el absurdo de la conclusión de que tiene mentalidad (aunque por supuesto mi argumento es vulnerable a la introducción de tal explicación). La cuestión, entonces, es si hay alguna buena razón para creer en el Funcionalismo. Un argumento a favor del Funcionalismo es que es la mejor solución disponible para el problema mente-cuerpo. Creo que éste es un mal argumento, pero puesto que también creo que el Psicofuncionalismo es preferible al Funcionalismo (por razones que mencionaré), pospondré la consideración de esta forma de argumentar hasta la discusión del Psicofuncionalismo.

El otro argumento que conozco a favor del Funcionalismo es que puede mostrarse que las identidades Funcionales son verdaderas sobre la base de los análisis de los significados de la terminología mental. De acuerdo con este argumento, las identidades Funcionales tienen que ser justificadas de la misma manera en que uno podría tratar de justificar la afirmación de que el estado de ser soltero es idéntico al estado de ser un hombre no casado. Un argumento similar apela a las trivialidades del sentido común acerca de los estados mentales en lugar de apelar a verdades acerca del significado. Lewis dice que las caracterizaciones funcionales de los estados mentales pertenecen al ámbito de la "psicología de sentido común, a la ciencia *folk*, más que a la ciencia profesional" (Lewis, 1972, pág. 250). (Véase también Shoemaker, 1975 y Armstrong,

1968. Armstrong tergiversa la cuestión de la analiticidad. Véase Armstrong, 1968, págs. 84-5, y pág. 90.) Y luego insiste en que las caracterizaciones Funcionales "deberían incluir sólo trivialidades que entre nosotros constituyen conocimiento común: todos las conocen, todos saben que todos las conocen, y así sucesivamente" (Lewis, 1972, pág. 256). Me referiré fundamentalmente a la versión "trivial" del argumento. La versión de la analiticidad es vulnerable a las mismas consideraciones, así como a dudas quineanas acerca de la analiticidad...

Estoy dispuesto a conceder, a los efectos del argumento, que es posible definir cualquier término de estado mental según las trivialidades concernientes a otros términos de estado mental, a términos de *input* y a términos de *output*. Pero esto no me compromete con el tipo de definición de términos de estado mental en la cual toda la terminología mental ha sido eliminada *via* la Ramsificación o algún otro mecanismo. Es simplemente falaz suponer que si cada término mental es definible en términos de los otros (más *inputs* y *outputs*), entonces cada término mental es definible no-mentalísticamente. Para ver esto, consideremos el ejemplo dado con anterioridad. Simplifiquemos la cuestión, claro, ignorando los *inputs* y los *outputs*. Definamos dolor como la causa de molestia y molestia como el efecto del dolor. Quien estuviera tan equivocado como para aceptar esto, no precisa aceptar una definición de dolor como *la causa de algo*, o una definición de molestia como *el efecto de algo*. Lewis sostiene que es analítico que dolor sea el ocupante de un cierto rol causal. Aun si estuviera en lo correcto acerca de un rol causal, especificado en parte mentalísticamente, uno no puede concluir que es analítico que dolor sea el ocupante de cualquier rol causal, especificado no-mentalísticamente.

No veo ningún argumento razonable a favor del Funcionalismo que se base en trivialidades o en la analiticidad. Además, la concepción que basa el Funcionalismo en trivialidades conduce a dificultades en los casos en que las trivialidades no tienen nada que decir. Recuérdese el ejemplo de los cerebros que son removidos para limpiarlos y rejuvenecerlos, en el que las conexiones entre nuestro cerebro y nuestro cuerpo se mantienen mediante radiotransmisión mientras continuamos con nuestra vida habitual. El proceso lleva unos pocos días y cuando se completa, el cerebro es reinsertado en el cuerpo. Ocasionalmente puede ocurrir que el cuerpo de una persona se destruya a causa de un accidente mientras el cerebro es limpiado y rejuvenecido. Si estuviera conectado a órganos sensoriales de *input* (pero no a órganos de *output*) tal cerebro no exhibiría *ninguna* de las conexiones usuales de carácter tri-

vial entre la conducta y los conjuntos de *inputs* y de estados mentales. Si como parece plausible, tal cerebro pudiera tener casi los mismos estados mentales (en sentido estrecho) que nosotros tenemos (y dado que tal estado de cosas podría volverse típico), el Funcionalismo estaría equivocado.

Resulta instructivo comparar la manera en que el Psicofuncionalismo intenta lidiar con los cerebros en cubetas. De acuerdo con el Psicofuncionalismo, es una cuestión empírica lo que va a valer como *inputs* y *outputs* de un sistema. Considerar a los impulsos neurales como *inputs* y *outputs* evitaría los problemas esquematizados, puesto que los cerebros en cubetas y los paralíticos podrían tener los impulsos neurales correctos aun sin tener movimientos corporales. Objeción: podría darse una parálisis que afecte el sistema nervioso, y afecte, de este modo, a los impulsos neurales, así el problema que se le plantea al Funcionalismo se le plantea también al Psicofuncionalismo. Respuesta: las enfermedades del sistema nervioso pueden, en efecto, *cambiar la mentalidad*, por ejemplo pueden hacer que los pacientes sean incapaces de sentir dolor. De este modo, podría ser verdad que una enfermedad del sistema nervioso ampliamente extendida que causa parálisis intermitente, hiciera a la gente incapaz de tener ciertos estados mentales.

De acuerdo con las versiones plausibles del Psicofuncionalismo, la tarea de decidir qué procesos neurales contarían como *inputs* y como *outputs* es, en parte, una cuestión de decidir *qué disfunciones cuentan como cambios en la mentalidad y qué disfunciones cuentan como cambios en las conexiones de* input *y de* output *periféricas*. El Psicofuncionalismo cuenta con un recurso que el Funcionalismo no tiene, puesto que el Psicofuncionalismo nos permite *corregir la línea que trazamos entre dentro y fuera del organismo, de modo de evitar problemas del tipo que hemos discutido*. Todas las versiones del Funcionalismo yerran al intentar trazar esta línea sólo sobre la base del conocimiento del sentido común; las versiones "analíticas" del Funcionalismo yerran especialmente al intentar trazar la línea a priori.

2. *Psicofuncionalismo*

Al criticar el Funcionalismo apelé al siguiente principio: si una doctrina tiene una conclusión absurda para creer en la cual no hay una razón independiente, y si no hay manera de salvar el absurdo o de mostrar que es engañoso o irrelevante, y si no hay una buena razón para

creer en la doctrina que lleva directamente al absurdo, entonces no se acepte la doctrina. Dije que no había ninguna razón independiente para creer que la simulación funcional que apela a la cabeza-homuncular, tiene estados mentales. Sin embargo, *hay* una razón independiente para creer que la simulación *Psico*funcional que apela a la cabeza-homuncular tiene estados mentales, es decir, que una simulación Psicofuncional de uno sería Psicofuncionalmente equivalente a uno, de modo que toda teoría psicológica verdadera de uno sería también verdadera de la simulación. ¿Qué mejor razón podría haber para atribuirle estados mentales, cualesquiera que sean ellos, que estén dentro del dominio de la psicología?

Este punto muestra que cualquier simulación Psicofuncional de uno comparte nuestros estados mentales *no*-cualitativos. Sin embargo, en la próxima sección argumentaré que hay, no obstante, algunas dudas de que comparta nuestros estados mentales cualitativos.

2.1 ¿Son los *qualia* estados Psicofuncionales?

Comencé este artículo describiendo un dispositivo de cabeza-homuncular y sosteniendo que hay dudas, prima facie, acerca de que tenga estados mentales, especialmente de que tenga estados mentales cualitativos, como dolores, picazones y sensaciones de rojo. La duda especial acerca de los *qualia* puede ser explicada, quizá, pensando en los *qualia invertidos* más que en los *qualia ausentes*. Tiene sentido, o parece tenerlo, suponer que los objetos que dos personas llaman verdes, lucen a una de ellas de la manera en que lucen los objetos que ambas llaman rojos. Parece que podríamos ser funcionalmente equivalentes aun cuando la sensación que las fresas evocan en uno sea cualitativamente la misma que la sensación que el césped evoca en la otra. Imagínese una lente invertida que cuando se coloca en el ojo de un sujeto produce exclamaciones como "Las cosas rojas lucen ahora de la manera en que las cosas verdes acostumbraban lucir, y viceversa". Imagínese además, a un par de gemelos idénticos, a uno de los cuales se le han insertado las lentes al nacer. Los gemelos crecen normalmente, y a la edad de 21 años son funcionalmente equivalentes. Esta situación ofrece, al menos, alguna evidencia de que el espectro de cada uno está invertido con relación al del otro (véase Shoemaker, 1975, nota 17, para una descripción convincente de la inversión intrapersonal del espectro). Sin embargo, resulta difícil ver cómo dar sentido al análogo de la inversión del espectro con respecto a estados no-cualitativos. Imagínese un par de

personas, una de las cuales cree que p es verdadera y que q es falsa, mientras que la otra cree que q es verdadera y que p es falsa. ¿Podrían esas personas ser funcionalmente equivalentes? Resulta difícil ver cómo podrían serlo.[10] Ciertamente, resulta difícil ver cómo dos personas podrían tener sólo esa diferencia en las creencias y sin embargo que no existiese ninguna circunstancia posible en la cual esa diferencia en la creencia se revelara por sí misma en conductas diferentes. Los *qualia* parecen ser supervenientes [*supervenient*] a la organización funcional, de una manera como las creencias no lo son...

Existe otra razón para distinguir firmemente entre estados mentales cualitativos y no-cualitativos cuando hablamos de teorías funcionalistas:

10. Supongamos que un hombre que tiene una buena visión de los colores utiliza erróneamente 'rojo' para denotar verde y 'verde' para denotar rojo. Es decir que confunde las dos palabras. Dado que esta confusión es puramente lingüística, aunque diga de una cosa verde que es roja, no *cree* que sea roja, así como un extranjero que ha confundido 'estuche' con 'sandwich' no cree que la gente come estuches en el almuerzo. Digamos que la persona que ha confundido de esta manera 'rojo' y 'verde', es una víctima de Cambio de Palabras [*Switching Word*].

Considérese ahora una enfermedad diferente: tener lentes que invierten rojo/verde ubicadas en los ojos, sin saberlo. Digamos que una víctima de esta enfermedad es una víctima de Cambio de Estímulo [*Stimulus Switching*]. Como la víctima de Cambio de Palabra, la víctima de Cambio de Estímulo aplica 'rojo' a las cosas verdes y viceversa. Pero la víctima de Cambio de Estímulo *tiene* creencias falsas acerca del color. Si se le muestra una mancha verde dice *y cree* que es roja.

Supongamos ahora que una víctima de Cambio de Estímulo de pronto se vuelve también una víctima de Cambio de Palabra (supongamos además que es un residente nativo de una villa remota del Ártico y que no posee creencias respecto de que el pasto sea verde, las fresas sean rojas, etcétera). Habla normalmente, aplicando 'verde' a las manchas verdes y 'rojo' a las manchas rojas. Por cierto, es funcionalmente normal. Pero sus *creencias* son tan anormales como eran antes de que se tornara una víctima de Cambio de Palabra. Antes de confundir las palabras 'rojo' y 'verde', aplicaba 'rojo' a una mancha verde, y erróneamente creía que la mancha era roja. Ahora (correctamente) dice 'rojo', pero su creencia sigue siendo errónea.

Así, dos personas pueden ser funcionalmente las mismas, aunque tengan creencias incompatibles. En consecuencia, el problema de los *qualia* invertidos infecta tanto a las creencias como a los *qualia* (aunque, presumiblemente, sólo a las creencias cualitativas). Este hecho debe interesar no sólo a quienes sostienen teorías de identidad de estados funcionales referidas a creencias, sino también a quienes se sienten atraídos por las explicaciones al estilo de Harman acerca del significado como rol funcional. Nuestra doble víctima —de Cambio de Palabra y de Cambio de Estímulo— es un contraejemplo para tales explicaciones. Porque su palabra 'verde' juega el rol normal en su razonamiento e inferencia, pero dado que al decir de algo que "es verde" expresa su creencia de que es *rojo*, usa 'verde' con un significado anormal. Estoy en deuda con Sylvain Bromberger por la discusión de esta cuestión.

el Psicofuncionalismo evita los problemas que el Funcionalismo tiene con los estados no-cualitativos, por ejemplo, las actitudes proposicionales como creencias y deseos. Pero el Psicofuncionalismo puede ser tan poco capaz de lidiar con los estados cualitativos, como lo es el Funcionalismo. La razón es que los *qualia* pueden muy bien no caer en el dominio de la psicología.

Para ver esto, permítasenos tratar de imaginar cómo sería una realización de la psicología humana que apelara a la cabeza-homuncular. La teorización psicológica corriente parece estar dirigida a la descripción de las relaciones del flujo-de-información [*information-flow*] entre mecanismos psicológicos. El objetivo principal parece consistir en descomponer tales mecanismos en mecanismos psicológicos primitivos, "cajas negras", cuya estructura interna cae en el dominio de la fisiología más que en el dominio de la psicología. (Véanse Fodor, 1968, Dennett, 1975 y Cummins, 1975; se plantean objeciones interesantes en Nagel, 1969.) Por ejemplo, un mecanismo cuasi-primitivo podría aparear dos ítemes en un sistema representacional y determinar si son casos del mismo tipo. O los mecanismos primitivos podrían ser como los de un computador digital, por ejemplo podrían ser (a) *agregue 1 a un registro dado*, y (b) *substraiga 1 de un registro dado, o si el registro contiene 0, pase a la instrucción n* (*indicada*). (Estas operaciones pueden combinarse para realizar cualquier operación de un computador digital; véase Minsky, 1967, pág. 206). Considérese un computador cuyo código de lenguaje-de-máquina contiene sólo dos instrucciones que corresponden a (a) y a (b). Si se pregunta cómo multiplica o resuelve ecuaciones diferenciales o compone nóminas, puede que se le conteste mostrándole un programa expresado en términos de las dos instrucciones del lenguaje-de-máquina. Pero si se pregunta cómo agrega 1 a un registro dado, la respuesta apropiada se da mediante un diagrama de los circuitos [*wiring diagram*], no mediante un programa. La máquina está construida [*hardwired*] para agregar 1. Cuando la instrucción que corresponde a (a) aparece en un cierto registro, los contenidos del otro registro cambian "automáticamente" de una cierta manera. La estructura computacional de un computador está determinada por un conjunto de operaciones primitivas y por las maneras en que las operaciones no-primitivas se arman a partir de aquéllas. De este modo, no importa a la estructura computacional del computador si los mecanismos primitivos son realizados mediante circuitos de tubos, circuitos de transistores o de relés. Del mismo modo no importa a la psicología de un sistema mental si sus mecanismos primitivos se realizan en uno u otro mecanismo neu-

rológico. Llámese a un sistema una "realización de la psicología humana" si toda teoría psicológica verdadera de nosotros es verdadera de él. Considérese una realización de la psicología humana cuyas operaciones psicológicas primitivas son efectuadas por hombrecitos, de la manera como lo fueron las simulaciones que apelan a la cabeza-homuncular ya discutidas. Así, quizás un hombrecito produzca ítemes de una lista uno a uno, otro compare estos ítemes con otras representaciones para determinar si se aparean, etcétera.

Ahora bien, existen buenas razones para suponer que este sistema tiene algunos estados mentales. Las actitudes proposicionales son un ejemplo. Quizá, la teoría psicológica identificará recordar que P con haber "almacenado" ["*stored*"] un objeto de carácter oracional [*sentence like*] que exprese la proposición que P (Fodor, 1975). Entonces, si uno de los hombrecitos puso un cierto objeto de carácter oracional en "almacenamiento", podemos tener razón para considerar al sistema como recordando que P. Pero a menos que tener *qualia* sea tener cierto procesamiento de información (en el mejor de los casos, una propuesta discutible) no existe una razón teórica tal para considerar al sistema como teniendo *qualia*. En resumen, hay quizá tantas dudas acerca de los *qualia* de este sistema con cabeza-homuncular como las que hay acerca de los *qualia* de la simulación Funcional que apela a cabeza-homuncular, discutida previamente en este artículo.

Pero, *ex hypothesi*, cualquier teoría psicológica es verdadera del sistema que estamos discutiendo. *Así, cualquier duda acerca de que tenga* qualia *es una duda acerca de que los* qualia *caigan en el dominio de la psicología*.

Podría objetarse: "¡La clase de psicología que se tiene en mente es la psicología *cognitiva*, es decir, la psicología de los procesos de pensamiento, y no es de extrañar que los *qualia* no caigan en el dominio de la psicología *cognitiva*!". Pero yo *no* tengo en mente a la psicología cognitiva, y si suena de esa manera, es fácilmente explicable: nada de lo que sabemos acerca de los procesos psicológicos que subyacen a nuestra vida mental conciente tiene que ver con los *qualia*. Lo que suele pasar por "psicología" de la sensación o del dolor es, por ejemplo, (a) fisiología; (b) psicofísica (es decir, el estudio de las funciones matemáticas que relacionan las variables de estímulo con variables de sensación; por ejemplo, la intensidad del sonido como una función de la amplitud de las ondas sonoras), o (c) un conjunto heterogéneo de estudios descriptivos (véase Melzack, 1973, cap. 2). De ellos, sólo la psicofísica podría ser interpretada como ocupándose de los *qualia* per se. Y es obvio que

la psicofísica sólo toca el aspecto *funcional* de la sensación, no su carácter cualitativo. Los experimentos psicofísicos hechos con uno tendrían los mismos resultados que si se hicieran con cualquier sistema Psicofuncionalmente equivalente a uno, aun si tuviera *qualia* invertidos o ausentes. Si los resultados experimentales no cambian, sea que los sujetos experimentales tengan o no tengan *qualia* invertidos o ausentes, difícilmente pueda esperarse que echen luz sobre la naturaleza de los *qualia*.

Por cierto que basándonos en la clase de aparato conceptual del que ahora disponemos en psicología, no veo cómo la psicología en alguna de sus encarnaciones actuales *podría* explicar los *qualia*. No podemos concebir ahora cómo la psicología podría explicar los *qualia*, aunque *podemos* concebir cómo la psicología podría explicar creer, desear, esperar, etcétera (véase Fodor, 1975). Que algo sea considerado inconcebible no es una buena razón para pensar que sea imposible. Mañana podrían desarrollarse conceptos que hicieran concebible lo que ahora es inconcebible. Pero todo lo que tenemos para seguir adelante es lo que sabemos y, si nos basamos en lo que tenemos para seguir adelante, pareciera que los *qualia* no caen en el dominio de la psicología...

No es una objeción a la sugerencia de que los *qualia* no son entidades psicológicas, afirmar que los *qualia* sean el paradigma mismo de algo que cae en el dominio de la psicología. Como se ha señalado a menudo, qué cae en el dominio de una rama particular de la ciencia es, en parte, una cuestión empírica. La liquidez del agua no resulta ser explicable por la química, sino más bien por la física subatómica. Las ramas de la ciencia abarcan en todo momento un conjunto de fenómenos que pretenden explicar. Pero puede descubrirse que algún fenómeno que parecía central a una rama de la ciencia pertenece, realmente, al ámbito de una rama diferente...

El Argumento de los *Qualia* Ausentes explota la posibilidad de que el estado Funcional o Psicofuncional que los Funcionalistas o Psicofuncionalistas querrían identificar con el dolor, pueda ocurrir sin que ningún *quale* ocurra. También parece ser concebible que ocurra un *quale* sin que ocurra dolor. Ciertamente, hay hechos que prestan apoyo a este punto de vista. Luego de las lobotomías frontales, los pacientes informan, típicamente, que todavía tienen dolores, aunque los dolores no los molestan ya (Melzack, 1973, pág. 95). Esos pacientes exhiben todos los signos "sensoriales" de dolor (tal como reconocer la agudeza de un pinchazo), pero a menudo no tienen deseo, o tienen pocos deseos, de evitar los estímulos "dolorosos".

Un punto de vista sugerido por estas observaciones es que cada

dolor es, en realidad, un estado *compuesto* cuyos componentes son un *quale* y un estado Funcional o Psicofuncional.[11] O lo que equivale a la misma idea, cada dolor es un *quale* que desempeña un cierto rol Funcional o Psicofuncional. Si este punto de vista es correcto, ayuda a explicar cómo la gente puede haber creído en teorías tan diferentes acerca de la naturaleza del dolor y de otras sensaciones; han enfatizado un componente a expensas del otro. Quienes proponen el conductismo y el funcionalismo tuvieron en mente un componente; quienes proponen la definición ostensiva privada han tenido en mente al otro. Ambas aproximaciones yerran en tratar de dar una explicación de algo que tiene dos componentes de naturalezas completamente diferentes.

3. *Chauvinismo* vs. *liberalismo*

Resulta natural entender las teorías psicológicas a las que el Psicofuncionalismo refiere, como teorías de la psicología *humana*. Entendido de este modo, es imposible para el Psicofuncionalismo que un sistema tenga creencias, deseos, etcétera, excepto que las teorías psicológicas que son verdaderas de nosotros sean verdaderas de él. El Psicofuncionalismo (entendido de ese modo) estipula que la equivalencia Psicofuncional con nosotros es necesaria para lo mental [*mentality*].

Pero aun cuando la equivalencia Psicofuncional con nosotros sea una condición de nuestro *reconocimiento de lo mental*, ¿qué razón hay para pensar que sea una condición de lo mental en sí mismo? ¿No podría existir una amplia variedad de procesos psicológicos posibles que subyazgan a lo mental, de los cuales instanciamos sólo un tipo? Supongamos que nos encontramos con marcianos y descubrimos que ellos son, de manera aproximada, Funcionalmente (pero no Psicofuncionalmente) equivalentes a nosotros. Cuando llegamos a conocerlos descubrimos que son tan diferentes de nosotros como los humanos que conocemos. Desarrollamos vastas relaciones culturales y comerciales con ellos. Cada cual estudia la ciencia y los periódicos filosóficos del otro, asiste a las películas del otro, cada cual lee las novelas del otro, etcétera. Entonces los psicólogos marcianos y los terrestres comparan sus anotaciones, sólo para descubrir que en la psicología subyacente, marcianos y terrestres

11. El *quale* podría ser identificado con un estado físico-químico. Este punto de vista concordaría con una sugerencia hecha por Hilary Putnam a fines de los 60 en su seminario de filosofía de la mente. Véase también cap. 5 de Gunderson, 1971.

son muy diferentes. Pronto acuerdan que la diferencia puede describirse como sigue. Pensemos en los humanos y en los marcianos como si fuesen productos de un diseño conciente. En tal proyecto de diseño habrá varias opciones. Algunas capacidades pueden ser asignadas por el diseño [*built in*](innatas), otras pueden ser aprendidas. El cerebro puede ser diseñado para llevar a cabo tareas usando tanta capacidad de memoria como sea necesaria para minimizar el uso de la capacidad computacional, o, por otra parte, el diseñador podría preferir conservar espacio de memoria y contar principalmente con capacidad computacional. Las inferencias pueden ser llevadas a cabo por sistemas que utilicen pocos axiomas y muchas reglas de inferencia o, en cambio, pocas reglas y muchos axiomas. Imagínese, ahora, que lo que los psicólogos marcianos y terrestres descubren cuando comparan sus anotaciones es que los marcianos y los terrestres difieren como si fueran los productos finales de elecciones de diseño maximalmente diferentes (compatibles con la equivalencia Funcional aproximada en los adultos). ¿Deberíamos rechazar nuestro supuesto de que los marcianos pueden disfrutar de nuestras películas, creer en los resultados científicos aparentes, etcétera? ¿Deberían "rechazar" su "supuesto" de que nosotros "disfrutamos" sus novelas, "aprendemos" de sus libros de texto, etcétera? Quizá no he proporcionado información suficiente para responder a esta pregunta. Después de todo, puede haber muchas maneras de completar la descripción de las diferencias humano-marcianas respecto de las cuales sería razonable suponer que no hay, simplemente, hechos decisivos, o suponer aun que los marcianos no merecen adscripciones mentales. Pero seguramente hay muchas maneras de completar la descripción de la diferencia marciano-terráquea que he esquematizado, según la cual sería perfectamente evidente que aun cuando los marcianos se comportaran de manera diferente de nosotros de acuerdo con experimentos psicológicos sutiles, no obstante, piensan, desean, disfrutan, etcétera. Suponer lo contrario sería puro chauvinismo humano. (Recuérdese que las teorías son chauvinistas en tanto que *niegan* falsamente que los sistemas tengan propiedades mentales, y liberales, en tanto *adscriben* falsamente propiedades mentales.)

Una sugerencia obvia para salir de esta dificultad consiste en identificar a los estados mentales con estados Psicofuncionales, considerando al dominio de la psicología de modo que incluya a *todas las criaturas con mentalidad*, incluidos los marcianos. La sugerencia es que definamos "Psicofuncionalismo" en términos de una psicología "universal" o "intersistémica" ["*cross-system*"], en lugar de la psicología humana, tal

como supuse antes. La psicología universal, sin embargo, es una empresa sospechosa. Porque, ¿cómo hemos de decir nosotros qué sistemas deben ser incluidos en el *dominio* de la psicología universal? Una manera posible de decidir qué sistemas tienen mentalidad y así cuáles caen en el dominio de la psicología universal, sería usar alguna *otra* teoría desarrollada de lo mental tal como el conductismo o el Funcionalismo. Pero tal procedimiento sería al menos tan carente de justificación como la otra teoría usada. Además, si el Psicofuncionalismo tiene que presuponer alguna otra teoría de la mente, podríamos muy bien aceptar en su lugar a la otra teoría de la mente.

Quizá la psicología universal evitará este problema del "dominio", del mismo modo que otras ramas de la ciencia lo evitan o buscan evitarlo. Otras ramas de la ciencia comienzan con dominios tentativos, basadas en versiones intuitivas y precientíficas de los conceptos que ellas, se supone, explican. Luego, intentan desarrollar clases naturales [*natural kinds*] de una manera que permite la formulación de generalizaciones legaliformes que se aplican a todas o a la mayoría de las entidades de los dominios precientíficos. En el caso de la mayoría de las ramas de la ciencia —incluyendo las ciencias biológicas y sociales tales como la genética y la lingüística—, el dominio precientífico resultó ser adecuado para la articulación de generalizaciones legaliformes.

Ahora bien, podría ser que fuéramos capaces de desarrollar una psicología universal de la misma manera en que desarrollamos la psicología terrestre. Decidimos, apoyados en una base intuitiva y precientífica, qué criaturas incluir en su dominio, y trabajamos para desarrollar clases naturales de la teoría psicológica, que aplicamos a todas o al menos a la mayoría de ellas. Quizás el estudio de una clase amplia de organismos que se encuentren en mundos diferentes conducirá un día a desarrollar teorías que determinen condiciones de verdad para la adscripción de estados mentales como creencia, deseo, etcétera, aplicables a sistemas que son preteóricamente diferentes de nosotros. Por cierto que tal psicología inter-mundos requerirá, sin duda, una clase completamente diferente de conceptos mentales. Quizás, habrá familias de conceptos que se correspondan con creencia, deseo, etcétera, es decir, una familia de conceptos similares a creencia [*belief-like concepts*], conceptos similares a deseo [*desire-like concepts*], etcétera. Si tal fuera el caso, la psicología universal que desarrollemos, sin duda dependerá, de alguna manera, de qué nuevos organismos descubramos primero. Aun cuando la psicología universal fuera de hecho posible, sin embargo, ciertamente habrá muchos organismos posibles cuyo *status* mental sea indeterminado.

Por otra parte, puede ser que la psicología universal *no* sea posible. Quizá la vida en el universo sea tal que, sencillamente, no tengamos bases para decisiones razonables acerca de cuáles son los sistemas que entran en el dominio de la psicología y cuáles no.

Si la psicología universal *fuera* posible, el problema que he estado planteando se desvanece. El Psicofuncionalismo-universal evita el liberalismo del funcionalismo y el chauvinismo del Psicofuncionalismo-humano. Pero la pregunta de si es posible la psicología universal es un interrogante que ahora no tenemos manera de responder.

He aquí una síntesis de lo argumentado:

1. El Funcionalismo tiene la consecuencia extraña de que la simulación de uno, apelando a la cabeza-homuncular, tiene *qualia*. Esto pone la carga de la prueba en el Funcionalista, a fin de que nos dé alguna razón para creer en su doctrina. Sin embargo, el único argumento que hay en la literatura a favor del Funcionalismo no es bueno, y así el Funcionalismo no da señales de satisfacer la carga de la prueba.
2. Las simulaciones Psicofuncionales que se hacen de nosotros comparten los estados mentales que caen en el dominio de la psicología, de modo que la cabeza-homuncular Psicofuncional no arroja duda sobre las teorías Psicofuncionales de los estados cognitivos, sino sólo sobre las teorías Psicofuncionales de los *qualia*, quedando la duda de si los *qualia* caen en el dominio de la psicología.
3. Las teorías Psicofuncionalistas de los estados mentales que caen en el dominio de la psicología son, sin embargo, irremediablemente chauvinistas.

Así, una versión del funcionalismo tiene problemas con el liberalismo y la otra con el chauvinismo. Porque en lo que respecta a los *qualia*, si caen en el dominio de la psicología, entonces el Psicofuncionalismo es a los *qualia* tan chauvinista como el Psicofuncionalismo lo es a la creencia. Por otra parte, si los *qualia* no caen en el dominio de la psicología, la cabeza-homuncular Psicofuncionalista puede ser usada contra el Psicofuncionalismo respecto de los *qualia*. Porque lo único que protege al Psicofuncionalismo del argumento de la cabeza-homuncular con respecto al estado mental S, es que si uno tiene S entonces cualquier simulación Psicofuncional de uno tiene que tener S; porque la teoría correcta de S se aplica tanto a ella como a uno.

3.1 El problema de los inputs y de los outputs

He estado suponiendo (como a menudo lo hacen los Psicofuncionalistas, véase Putnam, 1967) que los *inputs* y los *outputs* pueden especificarse mediante descripciones de impulsos neurales. Pero esta es una afirmación chauvinista, puesto que impide que organismos sin neuronas (por ejemplo, máquinas) tengan descripciones funcionales. ¿Cómo puede uno evitar el chauvinismo con respecto a la especificación de *inputs* y de *outputs*? Una manera sería caracterizar los *inputs* y los *outputs sólo como inputs* y *outputs*. Así, la descripción funcional de una persona podría listar *outputs* numerándolos: $output_1$, $output_2$... Entonces un sistema podría ser funcionalmente equivalente a uno si tuviera un conjunto de estados, *inputs* y *outputs* causalmente relacionados entre sí de la manera en que lo están los nuestros, cualesquiera sean los estados, *inputs* y *outputs*. Por cierto que aunque este enfoque viola la exigencia de algunos funcionalistas de que los *inputs* y los *outputs* sean especificados físicamente, otros funcionalistas —aquellos que insisten en que sólo las descripciones de *input* y de *output* sean *no-mentales*— pueden haber tenido en mente algo como eso. Esta versión del funcionalismo no "liga" [*tack down*] las descripciones relativamente específicas de los *inputs* y de los *outputs* a las descripciones funcionales en la periferia; más bien, esta versión del funcionalismo trata a los *inputs* y a los *outputs* como todas las versiones del funcionalismo tratan a los estados internos. Es decir, esta versión especifica estados, *inputs* y *outputs* requiriendo sólo que *sean* estados, *inputs* y *outputs*.

El problema con esta versión del funcionalismo es que es extremadamente liberal. Los sistemas económicos tienen *inputs* y *outputs*, tal como la entrada y salida de créditos y débitos. Y los sistemas económicos tienen también una rica variedad de estados internos, tal como tener una tasa de incremento del PBN igual al doble de la tasa mínima de interés. No parece imposible que un jeque acaudalado pudiera ganar el control de la economía de un país pequeño, por ejemplo Bolivia, y manipular su sistema financiero para hacerlo funcionalmente equivalente a una persona, por ejemplo a él mismo. Si esto parece implausible, recuérdese que los estados, *inputs* y *outputs* económicos que el jeque hace corresponder a sus estados, *inputs* y *outputs* mentales, no precisan ser magnitudes económicas "naturales". Nuestro jeque hipotético podría tomar *cualesquiera* de las magnitudes económicas, por ejemplo la quinta derivada del balance de pago. Su única limitación es que las magnitudes que elija sean económicas, que las magnitudes tengan tales

y cuales valores, sean *inputs*, *outputs* y estados, y que él sea capaz de montar una estructura financiera que se adecue al modelo formal propuesto. El mapeado [*mapping*] de las magnitudes psicológicas en las magnitudes económicas podría ser tan estrafalario como el jeque quiera.

Esta versión del funcionalismo es demasiado liberal y en consecuencia tiene que ser rechazada. Si hay puntos acordados cuando se discute el problema mente-cuerpo, uno de ellos es que la economía de Bolivia no podría tener estados mentales, no importa cuánto sea distorsionada por aficionados poderosos. Obviamente, tenemos que ser más específicos en nuestras descripciones de *inputs* y de *outputs*. La pregunta es: ¿existe una descripción de *inputs* y de *outputs* suficientemente específica como para evitar el liberalismo y, sin embargo, lo suficientemente general como para evitar el chauvinismo? Dudo que la haya.

Toda propuesta para la descripción de *inputs* y de *outputs* que he visto o pensado peca o bien de liberalismo o de chauvinismo. Aunque este trabajo se ha concentrado en el liberalismo, el chauvinismo es el problema más extendido. Considérense las descripciones Funcionales y Psicofuncionales estándar. Los Funcionalistas tienden a especificar los *inputs* y los *outputs* a la manera de los conductistas: los *outputs* en términos de movimientos de brazos y piernas, sonidos emitidos y cosas similares; los *inputs* en términos de luz y sonido penetrando en ojos y oídos... Tales descripciones son descaradamente *específicas-de-la-especie* [*species-specific*]. Los humanos tienen brazos y piernas, pero las víboras no, y sea que las víboras tengan o no tengan mentalidad, se puede imaginar fácilmente criaturas similares a las víboras que la tengan. Ciertamente, se puede imaginar criaturas con todo tipo de dispositivos *input-output*, por ejemplo, criaturas que se comunican y manipulan [cosas] mediante la emisión de fuertes campos magnéticos. Por supuesto, se podrían formular descripciones Funcionales para cada una de tales especies y en algún lugar del paraíso disyuntivo existe una descripción disyuntiva que abarcará a todas las especies que existen efectivamente en el universo (la descripción puede ser infinitamente extensa). Pero ni siquiera la apelación a entidades sospechosas tales como las disyunciones infinitas liberará al Funcionalismo, pues ni aun el punto de vista corregido nos dirá qué hay en común en los organismos que sienten dolor. Y no permitirá la adscripción de dolor a algunas criaturas hipotéticas (pero inexistentes) que sientan dolor. Más aún, éstas constituyen las bases sobre las cuales los funcionalistas rechazan, acerbamente, a las teorías disyuntivas propuestas a veces con desesperación por los fisicalistas. Si de pronto los funcionalistas vieran con agrado a los estados

como disyuntivas extravagantes para salvarse a sí mismos del chauvinismo, no tendrían manera de defenderse del fisicalismo.

Las descripciones Psicofuncionales estándar de *inputs* y de *outputs* son también específicas-de-la-especie (por ejemplo, en términos de actividad neural) y en consecuencia son también chauvinistas.

No resulta difícil explicar el chauvinismo de las descripciones *input-output* estándar. La variedad de vida inteligente posible es enorme. Dadas descripciones adecuadamente específicas de *inputs* y *outputs*, cualquier aprendiz de ciencia ficción en la edad secundaria será capaz de describir un ser sintiente y sapiente cuyos *inputs* y *outputs* no satisfagan esa descripción.

Argumentaré que *cualquier descripción física* de *inputs* y de *outputs* (recuérdese que muchos funcionalistas han insistido en las descripciones físicas) produce una versión del funcionalismo que inevitablemente es chauvinista o liberal. Imaginémonos con quemaduras tan graves que la manera óptima de comunicarnos con el mundo externo sea vía modulaciones de nuestro patrón personal EEG en Código Morse. Descubrimos que pensar un pensamiento excitante produce un patrón que la audiencia acuerda en interpretar como un punto y un pensamiento triste produce una "raya". Por cierto que esta fantasía no está lejos de la realidad. Según un artículo periodístico aparecido en el *Boston Globe* el 21 de marzo de 1976, "En UCLA, los científicos están trabajando en el uso del EEG para controlar máquinas... Un sujeto pone electrodos en su cuero cabelludo, y piensa un objeto a través de un laberinto". Presumiblemente, el proceso "inverso" es también posible: otros se comunican con uno en Código Morse mediante la producción de una descarga de actividad eléctrica que afecta nuestro cerebro (causando, por ejemplo, una posimagen [*afterimage*] extensa o breve). Recíprocamente, si los cerebroscopios que los filósofos a menudo imaginan se tornaran una realidad, nuestros pensamientos se leerían directamente de su cerebro. Nuevamente, el proceso inverso también parece posible. En estos casos, *el cerebro mismo se vuelve una parte esencial de los dispositivos* input *y* output *de uno.* Esta posibilidad tiene consecuencias embarazosas para los funcionalistas. Se recordará que los funcionalistas sostienen que el fisicalismo es falso porque un estado mental único puede ser realizado por una variedad indefinidamente grande de estados físicos que no tienen caracterización física necesaria ni suficiente. Pero si este argumento funcionalista contra el fisicalismo es correcto, *el mismo argumento vale para los* inputs *y los outputs*, puesto que la realización física de los estados mentales puede servir como una parte esencial de los dispositivos

de *inputs* y de *outputs*. Es decir, en cualquier sentido de 'físico' en el que la crítica funcionalista al fisicalismo es correcta, *no existirá caracterización física que valga para todos los sistemas mentales de inputs y de outputs, y sólo para ellos*. En consecuencia, todo intento de formular una descripción funcional con caracterizaciones físicas de *inputs* y *outputs*, excluirá inevitablemente algunos sistemas con mentalidad o incluirá algunos sistemas con mentalidad... *los funcionalistas no pueden evitar ni el chauvinismo ni el liberalismo*.

Así, las especificaciones físicas de *inputs* y de *outputs* no servirán. Más aún, la terminología mental o de "acción" (tal como "golpear a la víctima") tampoco puede usarse, puesto que usar tales especificaciones de *inputs* o *outputs* sería dejar a un lado al programa funcionalista que caracteriza a la mentalidad en términos no-mentales. Por otra parte, como se recordará, el caracterizar a *inputs* y *outputs* simplemente *como inputs* y *outputs*, es inevitablemente liberal. No veo cómo pueda haber un vocabulario para describir *inputs* y *outputs* que evite al liberalismo y al chauvinismo. No pretendo que este argumento sea concluyente contra el funcionalismo. Más bien, como el argumento funcionalista contra el fisicalismo, es mejor interpretarlo como un argumento acerca de la carga de la prueba. El funcionalista dice al fisicalista: "Es muy difícil ver cómo podría existir una única caracterización física de los estados internos de todas las criaturas con mentalidad, y sólo de ellas". Yo le digo al funcionalista: "Es muy difícil ver cómo podría existir una única caracterización física de los *inputs* y *outputs* de todas las criaturas con mentalidad, y sólo de ellas". En ambos casos, se ha dicho suficiente como para que el esbozo de cómo podrían ser posibles tales caracterizaciones sea responsabilidad de quienes piensan que podría haberlas.[12]

TRADUCTORA: Eleonora Baringoltz.
REVISION TÉCNICA: Eduardo Rabossi.

12. Estoy en deuda con Sylvain Bromberger, Hartry Field, Jerry Fodor, David Hills, Paul Horwich, Bill Lycan, Georges Rey y David Rosenthal, por sus comentarios detallados acerca de alguna de las primeras versiones de este trabajo. Partes de las primeras versiones fueron leídas a comienzos del otoño de 1975, en Tufts University, Princeton University, University of North Carolina en Greensboro, y State University of New York en Binghamton.

REFERENCIAS BIBLIOGRÁFICAS

Armstrong, D.: (1968) *A Materialist Theory of Mind*. Londres: Routledge & Kegan Paul.

Bever, T.: (1970) "The cognitive basis for linguistic structures", en J.R. Hayes (comp.), *Cognition and the Development of Language*. Nueva York, Wiley.

Block, N.: (1980) "Are absent qualia impossible?", *Philosophical Review* 89 (2).

Block, N. y Fodor J.: (1972) "What psychological states are not", *Philosophical Review* 81, 159-81.

Chisholm, Roderick: (1957) *Perceiving*. Ithaca, Cornell University Press.

Cummins, R.: (1975) "Functional analysis", *Journal of Philosophy* 72, 741-64.

Davidson, D.: (1970) "Mental events", en L. Swanson y J.W. Foster (comps.), *Experience and Theory*. Amherst, University of Massachusetts Press.

Dennett, D.: (1969) *Content and Consciousness*. Londres, Routledge & Kegan Paul.

Dennett, D.: (1975) "Why the law of effect won't go away", *Journal for the Theory of Social Behavior* 5, 169-87.

Dennett, D.: (1978a) "Why a computer can't feel pain", *Synthèse* 38, 3.

Dennett, D.: (1978b) *Brainstorms*, Montgomery, Vt., Bradford.

Feldman, F.: (1973) "Kripke's argument against materialism", *Philosophical Studies*, 416-19.

Fodor, J.: (1965) "Explanations in psychology", en M. Black (comp.), *Philosophy in America*, Londres, Routledge & Kegan Paul.

Fodor, J.: (1968) "The appeal to tacit knowledge in psychological explanation", *Journal of Philosophy* 65, 627-40.

Fodor, J.: (1974) "Special sciences", *Synthèse* 28, 97-115.

Fodor, J. y Garrett, M.: (1967) "Some syntactic determinants of sentential complexity", *Perception and Psychophysics* 2, 289-96.

Geach, P.: (1957) *Mental Acts*, Londres, Routledge & Kegan Paul.

Gendron, B.: (1971) "On the relation of neurological and psychological theories: A critique of the hardware thesis", en R.C. Buck y R.S. Cohen (comps.), *Boston Studies in the Philosophy of Science VIII*. Dordrecht, Reidel.

Grice, H.P.: (1975) "Method in philosophical psychology (from the banal to the bizarre)", *Proceedings and Addresses of the American Philosophical Association*.

Gunderson, K.: (1971) *Mentality and Machines,* Garden City, Doubleday Anchor.

Harman, G.: (1973) *Thought,* Princeton, Princeton University Press.

Hempel, C.: (1970) "Reduction: Ontological and linguistic facets", en S. Morgenbesser, P. Suppes y White (comps.), *Essays in Honor of Ernst Nagel.* Nueva York, St. Martin's Press.

Kalke, W.: (1969) "What is wrong with Fodor and Putnam's functionalism?", *Noûs* 3, 83-93.

Kim, J.: (1972) "Phenomenal properties, psychophysical laws, and the identity theory", *The Monist* 56 (2), 177-92.

Lewis, D.: (1972) "Psychophysical and theoretical identifications", *Australasian Journal of Philosophy* 50 (3), 249-58.

Locke, D.: (1968) *Myself and Others,* Oxford, Oxford University Press.

Melzack, R.: (1973) *The Puzzle of Pain,* Nueva York, Basic Books.

Minsky, M.: (1967) *Computation.* Englewood Cliffs, NJ, Prentice-Hall.

Mucciolo, L.F.: (1974) "The identity thesis and neuropsychology", *Noûs* 8, 327-42.

Nagel, T.: (1969) "The boundaries of inner space", *Journal of Philosophy* 66, 452-8.

Nagel, T.: (1970) "Armstrong on the mind", *Philosophical Review* 79, 394-403.

Nagel, T.: (1972) "Review of Dennett's *Content and Consciousness*", *Journal of Philosophy* 50, 220-34.

Nagel, T.: (1974) "What is it like to be a bat?", *Philosophical Review* 83, 435-50.

Nelson, R.J.: (1969) "Behaviorism is false", *Journal of Philosophy* 66, 417-52.

Nelson, R.J.: (1975) "Behaviorism, finite automata and stimulus response theory", *Theory and Decision,* 6, 249-67.

Nelson, R.J.: (1976) "Mechanism, functionalism, and the identity theory", *Journal of Philosophy* 73, 365-86.

Oppenheim, P. y Putnam, H.: (1958) "Unity of science as a working hypothesis", en H. Feigl, M. Scriven y G. Maxwell (comps.), *Minnesota Studies in the Philosophy of Science II,* Minneapolis, University of Minnesota Press.

Pitcher, G.: (1971) *A Theory of Perception,* Princeton, Princeton University Press.

Putnam, H.: (1963) "Brains and behavior"; reimpreso, como todos los artículos de Putnam citados en este trabajo (excepto "On proper-

ties"), en *Mind, Language and Reality, Philosophical Papers*, vol. 2, Londres, Cambridge University Press, 1975.

Putnam, H.: (1966) "The mental life of some machines".

Putnam, H.: (1967) "The nature of mental states" (publicado originalmente con el título de "Psychological Predicates").

Putnam, H.: (1970) "On properties", en *Mathematics, Matter and Method: Philosophical Papers*, vol. 1, Londres, Cambridge University Press.

Putnam, H.: (1975a) "Philosophy and our mental life".

Putnam, H.: (1975b) "The meaning of 'meaning' ".

Rorty, R.: (1972) "Functionalism, machines and incorrigibility", *Journal of Philosophy* 69, 203-20.

Scriven, M.: (1966) *Primary Philosophy*, Nueva York, Mc Graw-Hill.

Sellars, W.: (1956) "Empiricism and the philosophy of mind", en H. Feigl y M. Scriven (comps.), *Minnesota Studies in Philosophy of Science I*, Minneapolis, University of Minnesota Press.

Sellars, W.: (1968) *Science and Metaphysics* (cap. 6), Londres, Routledge & Kegan Paul.

Shoemaker, S.: (1975) "Functionalism and *qualia*", *Philosophical Studies* 27, 271-315.

Shoemaker, S.: (1976) "Embodiment and behavior", en A. Rorty (comp.), *The Identities of Persons*, Berkeley, University of California Press.

Shallice, T.: (1972) "Dual functions of consciousness", *Psychological Review* 79, 383-93.

Smart, J.J.C.: (1971) "Reports of immediate experience", *Synthèse* 22, 346-59.

Wiggins, D.: (1975) "Identity, designation, essentialism, and physicalism", *Philosophia* 5, 1-30.

CAPÍTULO 5

LA CONTINUIDAD DE NIVELES EN LA NATURALEZA *

William Lycan

El Funcionalismo Contemporáneo en la filosofía de la mente comenzó con una distinción entre las nociones de *rol* [*role*] y *ocupante* [*occupant*]. Como es sabido, la seductora comparación de las personas (o de sus cerebros) con los computadores, dirigió nuestra atención a la distinción entre un programa de computación (considerado en abstracto) y el material [*stuff*] específico del que la computadora está físicamente hecha, que *realiza* el programa. Es la primera distinción, no la última, la que nos interesa *vis-à-vis* la interpretación, explicación, predicción y empleo del "comportamiento" de la máquina; la gente construye computadores para operar programas, y emplea el material físico que mejor se preste a esa tarea.

La distinción entre "programa" y "material que lo realiza" [*realizing stuff*], más familiarmente, entre "*software*" y "*hardware*", se adecuó de manera natural a la filosofía de la mente cuando Putnam y Fodor expusieron las implicaciones chauvinistas de la Teoría de la Identidad. Lo que las "fibras-c" y similares hacen, podría haber sido hecho —ese rol podría haber sido ejecutado— por alguna estructura físico-química diferente. Y seguramente, si el mismo rol fuera realizado, si las mismas funciones fueran realizadas por una neuroquímica basada en el silicio en lugar del carbono, o si nuestras neuronas individuales fueran paulatinamente reemplazadas por prótesis electrónicas que hicieran la misma labor; en tal caso, intuitivamente, nuestra vida mental no se vería afectada. Lo que importa es la función, no el funcionario; el programa, no el material que lo realiza; el *software*, no el *hardware*; el rol, no el ocupante. Así nace el Funcionalismo y la distinción entre estados o propiedades "funcionales" y "estructurales" de un organismo.

* "The Continuity of Levels of Nature", extracto de los capítulos 4 y 5 de *Consciousness* (Bradford Books, 1987). Con autorización del autor y de Bradford Books.

El Funcionalismo es la única doctrina positiva en toda la filosofía por la cual estaría dispuesto a matar (aunque no esté autorizado a hacerlo).[1] Y considero (algunos dicen que obsesivamente) que la distinción "rol"/"ocupante" es fundamental para la metafísica. Pero sostengo que la *implementación* de esa distinción en la filosofía de la mente reciente es a la vez errónea y perniciosa. Y mi propósito en este capítulo es atacar las dicotomías "*software*"/"*hardware*" y "función"/"estructura", en sus formas filosóficas usuales, exhibir algunas de las confusiones sustantivas y corregir algunos de los errores que ha surgido de ellas.

La jerarquía

En términos generales, mi objeción es que hablar de "*software*"/"*hardware*" refuerza la idea de una Naturaleza bipartita dividida en dos niveles, aproximadamente, el psico-químico y el "funcional" o supra-organizacional (superviniente) [*supervenient*], como distinto de la realidad, que es una *jerarquía* múltiple de niveles naturales, cada uno de los cuales está demarcado por generalizaciones nomológicas y supervenientes a todos los niveles que les son inferiores en el continuo.[2] Véase a la Naturaleza organizada jerárquicamente de este modo, y la distinción "función"/"estructura" *se tornará relativa*: algo es un rol, como opuesto a un ocupante, un estado funcional como opuesto a lo que lo realiza, o viceversa, sólo *respecto de* [*modulo*] un nivel designado de la naturaleza. Ilustremos esto.

La fisiología y la microfisiología abundan en ejemplos. Las *células* —para tomar un término funcional mas bien conspicuo (!)— están constituidas por grupos cooperativos de ítemes más pequeños que incluyen membrana, núcleo, mitocondria, y demás: estos ítemes son en sí mismos *sistemas* de constituyentes cooperativos aun más pequeños. Además, aun los niveles más bajos de la naturaleza son numerosos y marcadamente distintos: el químico, el molecular, el atómico, el subatómico (tradicional), el microfísico. Los niveles son nexos de generalizaciones lega-

1. En ética, también creo firmemente en alguna forma de utilitarismo de acto, pero el sagrado principio de utilidad me prohíbe incluso decir esto, menos aún cometer crímenes (detectables) en su nombre.

2. La estructura jerárquica de múltiples niveles fue advertida y elocuentemente presentada por Herbert A. Simon (1969); no sé si la idea lo precedió. William C. Wimsatt escribió también brillantemente sobre ella (1976). Fodor (1968) y Dennett (1975) me llevaron a aplicarla a la psicología. Véanse referencias más adelante.

liformes interesantes, y se individuan de acuerdo con los tipos de generalización involucrados. Pero las células, si miramos hacia arriba en la jerarquía, se agrupan en tejidos, que se combinan para formar órganos, que se agrupan ellos mismos en sistemas de órganos, que cooperan —maravillosamente— para abarcar *organismos* completos, tales como seres humanos. Los organismos, además, se agrupan a su vez en grupos organizados (*organ-izados*). Y no hay clara diferencia de clase entre lo que consideramos corrientemente un organismo individual y grupos de organismos que funcionan corporativamente de una manera marcadamente unilateral [*single minded*]; "organismos grupales" en sí mismos, por así decir.[3]

A esta imagen agregativa abajo-arriba [*bottom-up*] de la organización jerárquica de la Naturaleza corresponde la familiar estrategia explicativa arriba-abajo [*top-down*].[4] Si queremos saber cómo se eliminan de los cuerpos humanos los desperdicios y toxinas, buscamos y hallamos un *sistema excretor* entrelazado con los sistemas digestivo y circulatorio. Si observamos al sistema con atención encontramos que trata a los residuos solubles en agua y no-solubles en agua, de modo diferente (lo cual no es sorprendente). Hallamos un *riñón*, que elabora en particular los desperdicios solubles. Si investigamos más en detalle, procediendo hacia abajo en la jerarquía de niveles, encontramos al riñón dividido en el córtex renal (un filtro) y la médula (un colector). El córtex está compuesto básicamente de nefrones. Cada nefrón tiene un glomérulo al que accede una arteriola aferente, y una cápsula muscular contráctil que controla la presión (la presión empuja el agua y los solutos a través de las paredes capilares dentro de la Cápsula de Bowman, dejando atrás las células sanguíneas y las proteínas sanguíneas mayores). La reabsorción y fenómenos similares se explican en términos de células, por ejemplo, por las propiedades especiales de las células epiteliales que limitan el largo túbulo del nefrón; estas propiedades especiales son a su vez explicadas en términos de la físico-química de las membranas celulares.

El cerebro no es una excepción a esta imagen jerárquica del orga-

3. Tengo en mente la discusión que hace Lewis Thomas (1974) de las sociedades de insectos y de la relación entre los seres humanos y sus mitocondrias. El análisis mereológico de los organismos depende mucho de los intereses teóricos. Nótese bien que podemos garantizar un pluralismo de diferentes relaciones reductivas entre niveles de la naturaleza. Considérese también la noción defendible de la corporación como persona (Biro, 1981; French, 1984; Brooks 1986).

4. Para una exposición y defensa lúcida de la estrategia, véase Cummins (1983). De cualquier modo, Richardson (1983) formula algunas críticas agudas.

nismo y de sus órganos. Las *neuronas* son células compuestas de *sómata* que contienen un núcleo y protoplasma, y fibras vinculadas a esos sómata que tienen una función efectivamente aislable; y se habla de ítemes funcionales más pequeños aún, tales como bombeadores iónicos que mantienen en el interior una alta concentración de potasio. Las neuronas mismas se agrupan en redes nerviosas y otras estructuras, tales como formaciones columnares, que a su vez se combinan para formar partes del cerebro, más grandes y más claramente funcionales (aunque no tan obviamente modulares). El sistema auditivo es un buen ejemplo. Hay evidencia de que el córtex auditivo presenta una organización columnar de dos dimensiones:[5] columnas de células diversamente especializadas distribuidas a lo largo de un axis responden selectivamente a frecuencias indicadas por impulsos que ingresan a través del nervio auditivo, mientras que columnas aproximadamente perpendiculares a éstas, de algún modo coordinan los *inputs* de un oído con los del otro. Las sensibilidades específicas de las células especializadas se explican, a su vez, por referencia a la transferencia de iones a través de las membranas celulares, y así hacia abajo. Por su parte, el córtex auditivo interactúa con otras agencias [*agencies*] de nivel superior —el tálamo, el colículo superior y otras áreas corticales, cuyas interacciones están altamente estructuradas.

Así colaboran una ontología agregativa y una epistemología arriba-abajo de la naturaleza. Esa colaboración ha sido elocuentemente defendida para la ciencia de la psicología, en particular por Attneave (1960), Fodor (1968), y Dennett (1975). Desarrollaré este punto con alguna extensión, siguiendo a Lycan (1981a).

Funcionalismo homuncular [homuncular functionalism]

Dennett (1975) se inspira en la metodología de ciertos proyectos de investigación de inteligencia artificial (IA).[6]

5. Para discusiones filosóficas relevantes y referencias véase P.M. Churchland (1986) y P.S. Churchland (1986).

6. El principal objetivo de Dennett en la obra que contiene el pasaje que sigue, es la explicación de la intencionalidad. Ese objetivo no es el mío aquí. Me interesa solamente el análisis de la teoría homuncular per se.

> En IA el investigador *comienza* con un problema caracterizado intencionalmente (por ejemplo, ¿cómo puedo lograr que un computador *entienda* preguntas del inglés?), lo descompone en subproblemas también caracterizados intencionalmente (por ejemplo, ¿cómo puedo lograr que una computadora *reconozca* preguntas, *distinga* sujetos de predicados, *ignore* análisis irrelevantes?) y luego descompone estos problemas más aún, hasta que finalmente alcanza descripciones de problemas o rutinas que son obviamente mecánicos.

Dennett extrapola este pasaje metodológico en el caso de la psicología humana y considero que sugiere que vemos a cada *persona* como una entidad corporativa [*corporate entity*], que corporativamente lleva a cabo muchas funciones, inmensamente complejas, de las usualmente llamadas mentales o psicológicas. Un psicólogo que adopte la metodología de Fodor y Dennett, inspirada en la IA, describirá a esa persona por medio de un diagrama de flujo [*flow chart*] que represente sus agencias subpersonales [*sub-personal agencies*] inmediatas y las variadas y diversas rutas de acceso mutuo que las habilitan para cooperar, al llevar a cabo los propósitos de la "institución" que las contiene o el organismo que cada persona es. Cada una de las agencias sub-personales inmediatas, representada en el diagrama de flujo por una "caja negra", es descriptible, a su vez, por su propio diagrama de flujo, que *la* descompone en agencias sub-sub-personales que cooperan para cumplir *sus* propósitos, y así sucesivamente. En esta concepción, las capacidades psicológicas de una persona y las diversas unidades administrativas de una organización corporativa mantienen jerarquías funcionales del mismo tipo y en un sentido similar.

Caracterizar los interrogantes de los psicólogos del modo como lo he hecho, equivale a verlos como admitiendo, en primer lugar, en el nivel de los datos o fenómenos, ciertas habilidades del ser humano caracterizadas intencionalmente o en términos de algún otro modo psicológico y, en segundo lugar, postulando —como entidades teóricas— los homúnculos [*homunculi*] o agencias sub-personales que se necesiten para explicar por qué los sujetos tienen esas habilidades. Luego los psicólogos postulan homúnculos más pequeños para explicar la conducta molar de los homúnculos previamente postulados, etcétera, etcétera. Es este rasgo del modelo Attneave/Fodor/Dennett el que bloquea con ingenio la objeción ryleana estándar del regreso al infinito en las teorías psicológicas homunculares.[7]

7. En realidad, como David Armstrong me ha señalado, la maniobra bloquea un número de argumentos típicos de regreso al infinito en la filosofía de la mente, incluyendo

Explicamos la actividad exitosa de un homúnculo, no suponiendo inútilmente un segundo homúnculo en su interior que ejecute exitosamente la actividad, sino postulando un *equipo* que consiste en varios homúnculos más pequeños, menos talentosos pero más especializados a nivel individual, detallando las maneras en que los miembros del equipo cooperan para producir su *output* corporativo.

Los psicólogos cognitivos y los psicólogos de la percepción tienen una idea razonablemente aceptable de la clase de agencias sub-personales que tendrían que postularse en los seres humanos para explicar su habilidad para realizar las acciones y las demás funciones que realizan. Dennett (1978, cap. 9) menciona, en el nivel inmediatamente subpersonal, un "componente de salida impresa" ["*print-out component*"] o centro del habla [*speech center*],[8] un "componente ejecutivo superior o componente *de Control*" ["*higher executive or Control component*"], un "almacén de memoria a corto plazo o memoria de tope" ["*short term memory store or buffer memory*"], un "componente de análisis perceptual" ["*perceptual analysis component*"], y un "componente de resolución de problemas" ["*problem-solving component*"]. Y Dennett (1978, cap. 11) examina, con cierto detallismo clínico, una estructura sub-personal de múltiples niveles que modeliza la conducta que manifiesta dolor. "Conducta" debe entenderse aquí en un sentido muy rico, dado que Dennett toma en cuenta escrupulosamente no sólo las clases usuales de conductas que son moneda corriente entre los Conductistas filosóficos y los apóstoles de la psicología de sentido común, sino también fenómenos más sutiles: diferencias muy pequeñas en nuestras descripciones fenomenológicas del dolor; fenómenos raramente señalados, tales como el hiato temporal que se siente entre experimentar que nos quemamos y sentir el dolor profundo de la quemadura, y (lo que es más interesante desde el punto de vista homuncular) la notable variedad de efectos de las más diferentes clases de anestésicos y drogas en la vida de un paciente y en sus informes retrospectivos acerca del dolor. Consideraciones de esta índole sirven a los psicólogos (y a Dennett) como indicadores vívidos de las complejidades existentes en la organización funcional relevante del sistema nervioso central (SNC), al indicar los

---

la crítica de Ryle en contra de las teorías volitivas de la decisión. Dennett mismo la emplea en contra del "problema de Hume" acerca de las representaciones del conocimiento de sí mismo (1978, pág. 122 y siguientes).

8. Para una presentación homuncular real del centro del habla, véase la figura 1, pág. 262 de Lycan (1984).

diversos componentes de caja negra en los distintos niveles de organización institucional que tenemos que representar en nuestros diagramas de flujo organizados jerárquicamente —las clases de receptores, inhibidores, filtros, servomecanismos [*damping*], disparadores y demás elementos que tenemos que postular—, y los tipos de caminos comparativamente diversos que vinculan estos componentes entre sí y con los componentes funcionales de sus propietarios, tales como analizadores perceptuales, almacenes de información y centros del habla.

El enfoque homuncular, teleológicamente interpretado, tiene muchas ventajas. Las referiré cuando haya dicho un poco más acerca de la teleología. Entre tanto pongo mis cartas sobre la mesa respecto de la forma general de una identificación-por-tipos [*type-identification*] de lo mental con lo no tan obviamente mental. Propongo identificar un estado mental tipo con la propiedad de tener tal y cual estado de cosas institucionalmente caracterizado, que se da en uno o más de los departamentos o subagencias [*subagencies*] homunculares apropiados. (Las subagencias son las que podrían representarse en un diagrama de flujo, asociadas con los respectivos agentes en diversos niveles de abstracción institucional.) Lo mismo vale para eventos, procesos y propiedades mentales. Tener (uno) un dolor de tipo T, podríamos decir, equivale a que el sub-...sub-homúnculo que realiza característicamente Φ [*Φ-er*], esté en un estado característico $S_T(\Phi)$; o equivale, para una actividad característica $A_T(\Phi)$, a transcurrir en ese sub-...sub-homúnculo.

Homúnculos y teleología

Se puede replicar que las caracterizaciones "Φ-er" y "$S_T(\Phi)$", en sí mismas, sólo están implícitamente definidas por el mapa teleológico del organismo, y que sus explicaciones contendrán, a su vez, referencias no eliminables a otras agencias y estados del organismo teleológicamente caracterizados. Esto es plausible, pero relativamente inofensivo. Nuestra labor como filósofos de la mente era explicar lo mental de un modo reductivo (y no circular), y esto es lo que estoy haciendo al reducir la caracterización mental a caracterizaciones homúnculo-institucionales, que son caracterizaciones teleológicas en diversos niveles de abstracción funcional. No estoy requiriendo, además, reducir las caracterizaciones institucionales a caracterizaciones más "finas" [*nicer*], más estructurales; si hubiera una reducción de tipos institucionales a, digamos, tipos fisiológicos, entonces, para el Homuncofuncionalismo [*Homunctionalism*],

la Teoría de la Identidad sería verdadera. Los *tipos* institucionales (en cualquier nivel jerárquico de abstracción dado) son irreducibles, aunque asumo del principio al fin, que los ejemplares [*tokens*] institucionales son reducibles, en el sentido de identidad estricta, al nivel subatómico.

En realidad, la irreducibilidad de los tipos institucionales es un dato a favor del Homuncofuncionalismo, como teoría filosófica de lo mental. Como observaron Donald Davidson y Wilfrid Sellars, una teoría adecuada de la mente tiene que explicar, entre otras cosas, la existencia misma del problema mente-cuerpo; esto supondría explicar por qué los estados mentales *parecen ser* tan diferentes de los físicos como para provocar un Cartesianismo ingenuo; por qué ha sido históricamente tan difícil, incluso para las versiones más sofisticadas, formular una reducción plausible de lo mental a lo físico, y por qué nuestra familia de conceptos mentales parece implicar una "totalidad homogénea" ["*seamless whole*"], conceptualmente no relacionada con la familia de conceptos fisiológicos o físicos.[9] El Homuncofuncionalismo proporciona los rudimentos de estas explicaciones. La aparente irreducibilidad de lo mental es la irreducibilidad genuina de los tipos institucionales a los menos teleológicos.[10] La dificultad de bosquejar una reducción defendible de lo mental incluso a lo institucional se debe a nuestra ignorancia de las operaciones organizativas de la institución misma en un nivel suficientemente bajo de abstracción. La irreducibilidad de tipos institucionales a tipos más fisiológicos no es un problema, en la medida en que nuestro sistema de categorías institucionales, nuestro sistema de categorías fisiológicas y nuestro sistema de categorías físicas, sean agrupamientos alternativos de los mismos casos.

Algunos filósofos podrían considerar poco atractiva la "reducción" Homuncofuncionalista. Ciertamente, aburriría a cualquiera que previamente concibiera la caracterización teleológica de las cosas *en términos de* ítemes mentales tales como deseos o intenciones. Por supuesto que, como implica la discusión anterior, no entiendo el discurso teleológico de este modo; más bien, considero que los tipos mentales forman una pequeña subclase de los tipos teleológicos que ocurren en su mayoría en un alto nivel de abstracción funcional. Pero si esto es así, entonces ¿cómo entiendo *yo* lo teleológico?

9. Davidson (1970) insiste enérgicamente en esto.

10. Así, el ejemplo que dio Smart de la lógica de los enunciados referentes a "nación" como diferente de la lógica de los enunciados referentes a "ciudadanos" puede ser más apropiado de lo que él imaginó.

Por mi parte, sobre este punto general tengo poco con que contribuir. Espero, y me inclino a creer, que las caracterizaciones teleológicas que requiere el Homuncofuncionalismo, puedan ser explicadas independientemente en términos evolutivos. Esta esperanza está alentada, considerablemente, por la obra de Karl Popper, William Wimsatt, Larry Wright, Karen Neander y otros filósofos de la biología.[11] No puedo mejorar sus discusiones técnicas. De cualquier modo, quiero señalar un punto teórico y ofrecer un ejemplo en su apoyo.

El punto teórico es que el carácter teleológico o teleologicidad [*teleologicalness*] de las caracterizaciones es una cuestión de grado: algunas caracterizaciones de algo son más teleológicas que otras. Una y la misma raja [*slice*] de espacio-tiempo puede ser ocupada por una colección de moléculas, un trozo de material muy duro, una barra de metal con una paleta, lo que desplaza al pasador de una cerradura, una llave, un instrumento para abrir una puerta, lo que permite la entrada a un cuarto de hotel, lo que facilita relaciones adúlteras, un destructor de almas. Así, no podemos dividir con nitidez nuestra teoría de la naturaleza en una parte "que se porta bien", puramente mecánica, y una parte vitalista desordenada y dudosa que es mejor ignorar o suprimir. Y por esta razón no podemos sostener que una reducción de lo mental a lo teleológico no dé beneficios en cuanto a ductilidad ontológica; las caracterizaciones altamente teleológicas, a diferencia de las caracterizaciones mentales ingenuas y explicadas, tienen la virtud de transformarse sin solución de continuidad en caracterizaciones (más) crasamente físicas.[12]

11. Popper (1972), Wimsatt (1972), Wright (1973), Millikan (1984), Neander (1981, 1983). La explicación evolucionista que Neander ofrece es la mejor que conozco. Ella ha sido criticada con eficacia por E. Prior en un trabajo inédito y por Pargetter y Bigelow (1987); me parece que la verdad está en algún punto intermedio. Jonathan Bennett (1976) ofrece un enfoque naturalista de la teleología fundado en Ann Wilbur MacKenzie (1972) (y en ocasiones me ha sugerido cambiar de posición).

12. Las caracterizaciones de los contenidos de nuestra raja de espacio-tiempo pueden así ser dispuestas en un continuo, desde las menos teleológicas hasta las más (altamente) teleológicas. Ese continuo corresponde claramente a la jerarquía de ejemplificaciones o realizaciones funcionales. Las moléculas realizan en conjunto o tienen el rol de la pieza de metal; la pieza de metal tiene el rol de la llave; la llave es nuestro instrumento para abrir la puerta, y así sucesivamente. Creo que la importancia de las jerarquías funcionales de este tipo es lo que alienta la reducción ontológica y la idea de que "en última instancia todo es físico". Para la relación entre la teleología vista desde una perspectiva evolucionista, las jerarquías funcionales, la ontología y la metodología de la reducción científica, véase nuevamente Wimsatt (1976). Me ha beneficiado también la lectura de Mellik (1973) y Matthen y Levy (1984).

Permítaseme dar un ejemplo relevante para la psicología. Consideremos un organismo capaz de *reconocer rostros* (para tomar uno de los mejores ejemplos de Dennett de una capacidad psicológica programable). Es perfectamente lícito preguntar *cómo* hace el organismo su labor; la criatura podría concretar su reconocimiento de rostros al ser construida de acuerdo con un número de planes funcionales enteramente distintos. Supongamos que el plan particular que se emplea es el siguiente: la criatura aceptará la orden de identificar, sólo cuando se dé como *input* una visión frontal, de perfil derecho o de perfil izquierdo. La rutina ejecutiva dirigirá al campo perceptual [*perceptual display*] el *localizador de perspectiva* [*view point locator*], que clasificará el *input* en una de las tres posibles categorías de orientación. El campo perceptual se exhibirá entonces al *analizador* [*analyzer*] apropiado, el que producirá como *output* una codificación del contenido del campo perceptual. Un *archivo* [*librarian*] confrontará esta fórmula codificada con el *stock* de informes visuales similarmente codificados, ya almacenados en la memoria del organismo; si encuentra un semejante, mirará el rótulo de identificación adscripto a la fórmula codificada del semejante, y mostrará el rótulo al *oficial de relaciones públicas* [*public relations officer*] del organismo, quien dará instrucciones fonológicas a las *subrutinas motoras* que harán que el organismo pronuncie públicamente y en voz alta un nombre.

Sabiendo que ésta es la manera en que nuestro reconocedor de rostros hace su trabajo, podríamos desear conocer más detalles. Podríamos desear conocer cómo trabaja el localizador de perspectiva (¿Es una simple pantalla?), cómo está organizada la oficina de relaciones públicas o qué clase de sub-componentes emplea el analizador. Supongamos que se descubre que el analizador consiste en un *proyector* que impone una grilla al campo visual y en un dispositivo de lectura [*scanner*] que recorre la grilla un cuadrado por vez y produce un número en código binario. Podemos ir más allá y preguntar cómo trabaja el dispositivo de lectura, y encontrar que consiste básicamente en un fotómetro que registra un grado de oscuridad en cada cuadrado e informa "0" o "1" según sea el caso; podemos preguntar cómo trabaja el fotómetro y recibir información acerca de sustancias químicas fotosensibles, etcétera, etcétera. Ahora bien, ¿en qué punto de este descenso a través de la jerarquía institucional (del *reconocedor de rostros* al *dispositivo de lectura* al *fotómetro* a la *substancia fotosensible*, y hasta donde uno quiera seguir) deja nuestra caracterización de ser teleológica y comienza a ser puramente mecánica? Creo que es claro que no hay tal punto, sino más bien un

continuo finamente tramado que conecta lo abstracto y altamente teleológico con lo granularmente concreto y sólo marginalmente teleológico. Y ésta es la razón por la cual lo mental puede *parecer* totalmente distinto y aislado de lo físico-químico sin *ser* ontológicamente tal cosa.[13]

Una palabra final sobre mi dependencia de la teleología, apenas explicada: no pretendo que esa teleología apenas explicada sea buena o deseable. Por mi parte, no me gusta. Mi tesis es que el misterio de lo mental *no es mas grande que* el misterio del corazón, el riñón, el carburador o la calculadora de bolsillo. Y como tesis ontológica es muy reconfortante.[14]

13. Como Jerry Fodor me ha señalado personalmente, hay una distinción tolerablemente clara que un teórico de los dos niveles podría tener en mente, y que es absoluta: la distinción entre objetos cuyas partes propias son esenciales y aquellos cuyas partes no lo son. Por ejemplo, las partes de una bicicleta o incluso de un árbol son reemplazables mientras que las partes de una molécula de agua tal vez no lo sean (podríamos argüir que si la molécula perdiera alguno de sus átomos de hidrógeno u oxígeno ya no sería esa molécula, o que sin los átomos que correspondan no sería en absoluto una molécula de agua). Pienso que la distinción es genuina y confío que tenga alguna importancia metafísica. Pero no tiene importancia *psicológica*. El nivel de los componentes químicos está demasiado lejos en la jerarquía institucional como para afectar el nivel mental; es decir que si dos neuroanatomías son exactamente iguales, aunque estén realizadas por diferentes componentes químicos, la psicología será la misma.

14. Amelie Rorty me sugirió la idea aristotélica de explicar la función de los componentes de un organismo (más exactamente, de explicar la constitución-para-el-éxito-de-sus-funciones [*its-functions'-constituting-its-thriving*]) por referencia a la adecuación de tales funciones a las condiciones materiales de la especie a la que el organismo pertenece. Esta idea calza con la noción etiológica de función que promuevo. Dada una masa relativamente indiferenciada de material biológico "inferior" en un nivel evolutivo muy primitivo, ¿cómo podría agruparse y articularse para afrontar el mundo plenamente de un modo más seguro y menos vulnerable? Su propia naturaleza "estructural" o "material" nos forzaría a algunas respuestas alternativas y sugeriría algunas otras, y dadas las presiones de la selección de varios tipos ahora retrodecibles, no es sorprendente que algunas o la mayoría de estas alternativas se hayan realizado. Si "función" se entiende en términos evolutivos, entonces, la función misma queda explicada de este modo en términos de las propensiones del sustrato material de un organismo. Considero que esta explicación complementa las de "causación descendente" basadas en niveles superiores de la naturaleza (de la clase que menciona Wimsatt), más bien que compite con ellas. En realidad, se tiene una especie de movimiento de pinza: una presión selectiva desde los niveles más altos, que interactúa con la presión ascendente que surge de la naturaleza y las propensiones de la constitución química particular del entorno preexistente. Ambas presiones moldean juntas lo que yace entre ellas. Pero uno podría querer enfatizar la presión ascendente a expensas de la explicación a partir de los niveles superiores. *En algún sentido* el énfasis tiene que ser correcto, dada la superveniencia del nivel último respecto del más bajo, aunque es difícil descubrir todas las diferentes interrelaciones que hay de arriba hacia abajo. Rorty me señaló (por carta) que la realizabilidad múltiple a plena escala tiene que ser dis-

Ventajas del enfoque teleológico

El lector habrá advertido que tomo *función* muy en serio y literalmente: como verdadera-y-honesta teleología natural.[15] La decisión de tomar "función" teleológicamente tiene ciertas virtudes clave: (i) como hemos visto, una interpretación teleológica de "función" ayuda a dar cuenta de la *homogeneidad* [*seamlessness*] con que es percibido lo mental, de la interrelación de los conceptos mentales entre sí de un modo que no parece tener ninguna relación con conceptos físicos y químicos.[16] (ii) La imposición de un requisito teleológico a la noción de realización funcional, nos permite superar los contraejemplos estándar al Funcionalismo de Máquina y, como argumentaré, a cualquiera otra clase de Funcionalismo; ver más adelante. (iii) Un funcionalismo teleológico nos ayuda también a entender la naturaleza de las *leyes* biológicas y psicológicas, particularmente frente al escepticismo davidsoniano acerca de estas últimas (Lycan, 1981b; Cummins, 1983). (iv) Si las caracterizaciones teleológicas son explicadas en sí mismas en términos evolutivos, entonces nuestras propias capacidades de tener estados mentales se hacen más fácilmente explicables en términos de causas finales. Resulta

---

tinguida de la mera caracterización funcional de los estados de un organismo, dado que el tratamiento detallado de la función tiende a imponer requerimientos estrictos sobre el material-en-que-se-realiza. Hay aquí un intercambio. Pero no veo que la estrategia de explicación aristotélica de-arriba-hacia-abajo esté per se en contra de la realizabilidad múltiple. Pues las mismas respuestas o soluciones podrían muy bien ser descubiertas a partir de trozos de materia prima absolutamente diferentes en su composición química. Rorty ofrece el ejemplo de *comer*. Los computadores no comen, en ningún sentido literal, y la Tierra no ingiere la lluvia; la realizabilidad múltiple fracasa incluso si la actividad está caracterizada funcionalmente. Quiero dar la misma clase de réplica que daré más adelante a un argumento de Block: por supuesto que los computadores y otras entidades (incluso biológicas) no comen, pero hay una caracterización intermedia y más abstracta del comer mismo —*holotropismo*, se lo llamaba en mi época de estudiante—, que excluye los computadores pero incluye una cantidad de especies bioquímicamente diferentes de la nuestra; ella tiene alguna relación con la adquisición de proteínas, muy similar a la homogeneización física, ingesta y asimilación que hacemos de ellas, sin mayor reagrupamiento de aminoácidos ni nada por el estilo; de cualquier modo es una forma de nutrición que se distingue netamente de las de otras especies y es distintiva de nuestro phylum, o el de cualquier otro. Este punto concuerda bien con mi idea de las caracterizaciones funcionales, que se hace cargo de los niveles intermedios de la naturaleza y que no es ni demasiado vaga y general ni demasiado limitada chauvinísticamente a una especie.

15. Elliot Sober (1985) elogia esta actitud como "devolviendo la función al funcionalismo"; véanse mis comentarios en la página 27 de Lycan (1981a) con referencia a la puja de Fodor y Putnam sobre la palabra "función".

16. Para un tratamiento detallado del tema, véase Lycan, 1981b.

más obvio por qué tenemos dolores, creencias, deseos y demás.[17] (v) El enfoque teleológico proporciona el punto de partida de un tratamiento de la *intencionalidad* que evita las dificultades estándar de otros tratamientos naturalistas y, en particular, permite que los estados y eventos cerebrales tengan un contenido intencional *falso*. Las teorías causales y nomológicas de la intencionalidad tienden a vacilar respecto de esta última cuestión (véase Lycan, 1989).

He argumentado más arriba que necesitamos una noción gradualista de la teleología o que, por lo menos, permita caracterizaciones con diferente grado de teleologicidad; también argumenté que ya tenemos esa noción, aunque sea difícil de explicar: recuérdense los ejemplos del reconocedor de rostros y de la llave. Los filósofos pueden diferir entre sí acerca del análisis correcto de esta noción gradualista de la teleología; por mi parte tiendo a ver la graduación como determinada por la docilidad para ser explicada por causas finales, donde explicación "por causa final" debe reconstruirse, a su vez, como una suerte de explicación evolutiva (aunque algunos detalles quedan por ser elaborados). Pero dos puntos centrales ya son claros: (i) al menos para los organismos simples, los grados de caracterización de teleologicidad se corresponden muy bien con los niveles de la naturaleza,[18] y (ii) no hay ningún lugar particular *ni* en el continuo de la teleologicidad, *ni* entre los varios niveles de la naturaleza, en el que sea normal introducir una cuña decisiva, tal que las descripciones de la naturaleza puedan dividirse prolijamente en un modo puramente mecanicista, educado, puramente "estructural" y un modo más dudoso, intencional y quizá vitalista: ciertamente, ningún lugar que correspondiera a alguna distinción intuitiva entre lo psicológico y lo meramente químico, pues hay mucha biología de por medio.

Ahora se ven mis propias tendencias pan-psiquistas o al menos panteleológicas. Muchos filósofos más "duros" las hallarán fantasiosas. En el mejor de los casos, estoy dispuesto a admitir que es difícil ver utilidad

17. ¿Por qué duele el dolor? ¿Por qué no podemos tener un sistema indicador del daño que instigue su reparación, que no sea molesto? La respuesta es simple: supongamos que tengo un sistema como la luz roja que indica el recalentamiento del motor del automóvil. Así como suelo ignorar irracionalmente la luz roja y esperar vagamente que se apague, ignoraría la luz indicadora personal si no tuviera un motivo urgente para hacer algo al respecto.

18. Robert Van Gulick me proveyó (por correspondencia) de algunos casos meteorológicos y geológicos en los que (aparentes) grados de teleologicidad no se corresponden con niveles de la naturaleza. Estos casos son muy pertinentes, pero tendré que posponer su tratamiento.

alguna en considerar, digamos, la descripción del nivel *atómico* como teleológica en algún grado (y por supuesto también yo en mis momentos de lucidez estoy dispuesto a admitirlo).[19] Ciertamente la explicación-por-causas-finales no persiste en los niveles inferiores. *Pero*, la caracterización inconfundiblemente teleológica (descripción que es obviamente teleológica en algún grado, por pequeño que sea) llega *tan lejos* como pueda ser relevante para la psicología (por ejemplo, más allá de la neuroanatomía). Y la distinción *rol/ocupante* se extiende aun más abajo. Así, la alardeada distinción "función"/"estructura", tal como ordinariamente es concebida por los filósofos, fracasa en elucidar la psicología humana en la que habita....

Todo lo que he dicho hasta aquí puede parecer insípido y obvio. Espero que lo parezca. Trato de llamar la atención sobre lo que considero una verdad de entrecasa acerca de la estructura del mundo físico, porque pienso que pasar por alto esa verdad, no atender a la naturaleza jerárquica de la Naturaleza, ha conducido a errores significativos acerca de la conciencia y de los *qualia*. En lo que resta discutiré esto brevemente.

Block (1978; 1981), Lycan (1987) y otros, han adelantado varios contraejemplos, destinados a mostrar que la posesión de una organización funcional, por más compleja que sea, es insuficiente para dar cabida a los estados cualitativos, fenoménicamente experimentados; probablemente los mejor conocidos y más discutidos sean los contraejemplos de la "cabeza homuncular" y la "población de China" que ofrece Block. Para refutar tales contraejemplos, el Funcionalista tendrá que exhibir algún requerimiento razonable que ellos no logren satisfacer, a pesar de que imitan de un modo u otro la organización funcional de una criatura real con sensaciones.

El Homuncofuncionalismo, entendido teleológicamente, los supera con facilidad. Porque ninguno de los sistemas imaginados en los contraejemplos está teleológicamente organizado de la manera correcta; la mayoría, incluso ni siquiera son organismos. (Véase Lycan, 1987, capítulos 3 y 5.)

Aun cuando esos casos paradójicos no refuten al Homuncofuncio-

19. Ned Block, quien está en completo desacuerdo conmigo en este punto, me dijo una vez (en diálogo personal): "Te dejo *neuronas y células* y demás como funcionales, pero cuando llegamos al *hidrógeno* y al *oxígeno*, cuando bajamos hasta el nivel de la *química*, ¡no hay nada funcional o teleológico! ¿O sí? ¿"Hidro-"*qué*? ¿"Oxi"-*qué*? (El punto es modesto, pero inmensamente gratificante.)

nalismo, quedan sin resolver algunos problemas de chauvinismo y liberalismo. Estén o no acertados Fodor y Block al sugerir que Putnam se acercó al conductismo al alejarse de la Teoría de la Identidad, el Funcionalista tiene ciertamente la responsabilidad de hallar un nivel de caracterización de los estados mentales que no sea ni tan abstracto o conductista como para excluir la posibilidad del espectro invertido [*inverted spectrum*], etcétera, ni tan específico y estructural como para caer en el chauvinismo. Block mismo llega a argumentar que este problema es insoluble.

Block, en particular, hace surgir el dilema, respecto de la caracterización de *inputs* y *outputs*. Comúnmente, *inputs* y *outputs* no pueden caracterizarse en términos neurológicos humanos; esto impediría, chauvinísticamente, que atribuyéramos descripciones mentales a máquinas, marcianos y otras criaturas que difieren biológicamente de nosotros, sin importar las credenciales convincentes que pudieran ofrecer en defensa de su capacidad de tener sensaciones. Por otra parte, los *inputs* y *outputs* no pueden ser caracterizados en términos puramente abstractos (esto es, meramente como "*inputs*" y "*outputs*"), dado que ello nos conduciría a la clase de ultraliberalismo que Block ha desacreditado mediante sus primeros ejemplos y también mediante los nuevos, como el de un sistema económico que tuviera *inputs* y *outputs* muy complejos y estados internos, pero que ciertamente no tendría características mentales. Tampoco podemos apelar a clases particulares de interacciones de la entidad capaz de tener sensaciones, con su entorno, *via inputs* y *outputs*, dado que en algunos pocos casos (paralíticos, cerebros *in vitro*, y otros) querríamos atribuir descripciones mentales a objetos que no pueden interactuar con su medio en modo alguno. Block (1978) concluye:

> ¿Existe una descripción de *inputs* y *outputs* suficientemente específica como para evitar el liberalismo, pero suficientemente general como para evitar el chauvinismo? Lo dudo.
> Cada propuesta que he conocido o concebido de una descripción de *inputs* y *outputs* es culpable de liberalismo o chauvinismo. Aunque este artículo trata del liberalismo, el chauvinismo es el problema más persistente.
> [...] *no habrá una caracterización física que se aplique a todos los* inputs *y* outputs *de los sistemas mentales*. Por lo tanto, cualquier intento de formular una descripción funcional con caracterizaciones físicas de *inputs* y *outputs* inevitablemente excluirá algún sistema [posible] con vida mental o bien incluirá algunos sistemas sin vida mental.
> [...] Por otra parte, como recordarán, la caracterización de los *inputs* y *outputs* simplemente como *inputs* y *outputs* es inevitablemente liberal. De

> todas formas, no veo cómo el Funcionalismo podría describir *inputs* y *outputs* sin toparse con el liberalismo o el chauvinismo, o abandonar el proyecto original de caracterizar la mente en términos no mentales. No afirmo que éste sea un argumento concluyente contra el Funcionalismo. Más bien, igual que el argumento Funcionalista en contra del fisicalismo, estaría posiblemente mejor construido como un argumento de la carga de la prueba.

No estoy seguro de cuán detallado sea el plan que Block exige aquí al Funcionalista, aunque estoy de acuerdo en que, en una interpretación benevolente de "carga de la prueba", el Funcionalista *tiene* la carga de la prueba respecto del desafío de Block. La pregunta es si esa carga es tan excesivamente pesada como Block parece suponer. Y creo que hay al menos tres factores que la aligeran considerablemente y nos dan alguna razón para ser optimistas.

En primer lugar, hay una línea de argumentación que ofrece al menos una tenue razón positiva o una motivación natural para pensar que el dilema del chauvinismo y del liberalismo (ya sea en relación con los *inputs* y *outputs*, o con los estados internos que el Funcionalista identifica con nuestros estados mentales) tiene solución. Ella comienza como un argumento resbaladizo. Block ha presentado el dilema de modo muy intransigente, al suponer que las únicas opciones que uno tiene son (a) caracterizar los *inputs* y *outputs* fisiológicamente, y ser un chauvinista, o (b) caracterizar los *inputs* y *outputs* de manera "puramente abstracta", y ser un liberal. Pero esta presentación estricta de las alternativas olvida el hecho...de que la abstracción es una cuestión de grado. La caracterización puramente fisiológica es un extremo, que se encuentra en el punto más bajo o "más estructural" del espectro; la caracterización "puramente abstracta" es el extremo opuesto, que se halla en el punto más elevado o "más funcional". Nótese que... hay caracterizaciones que son aún *más* "estructurales" de lo que lo son algunas caracterizaciones fisiológicas, como las microfísicas, respecto de las cuales las fisiológicas son "funcionales"; de modo similar, hay en realidad caracterizaciones más abstractas que las de "*input*" y "*output*" mismas, tales como "transferencia", "movimiento", o incluso "ocurrencia". Si es verdad, como parece serlo, que las caracterizaciones "puramente abstractas" y las caracterizaciones fisiológicas meramente se ubican cerca de los dos extremos de un contínuo de abstracción funcional, entonces es razonable esperar que exista algún nivel intermedio de abstracción que produciría caracterizaciones que excluyan la economía boliviana, la Galaxia Abnegoniana, la microbiología de Everglades, y similares, pero dejen

lugar a los seres humanos, los moluscos, los marcianos y los cerebros *in vitro*. La verdad está (como casi siempre) en algún lugar intermedio, dependiendo de qué aspecto de la vida mental nos interese, aunque no siempre nos interese el mismo punto intermedio. Esperemos y veamos cuáles serán los recursos disponibles en los distintos niveles intermedios.[20]

Recordemos además (ésta es mi segunda respuesta al desafío de Block) que nada nos fuerza a suponer que los diferentes tipos de estados mentales ocurran en un *mismo* nivel de abstracción funcional. Las clases de estados mentales intuitivamente "más conductuales", tales como las creencias, los deseos y las intenciones, presumiblemente ocurren en un nivel de abstracción relativamente elevado, y esto facilita que adscribamos creencias deseos e intenciones a los marcianos, cuya conducta pública y rasgos psicológicos superficiales sólo se asemejan a los nuestros de manera aproximada; lo mismo vale para las actividades mentales altamente "informacionales" tales como recordar y (literalmente) computar. Intuitivamente, los estados mentales más cualitativos, "menos conductales", probablemente ocurran en un nivel de abstracción mucho más bajo; los eventos sensoriales [*sensings*] con cierta clase especial de características cualitativas, probablemente *sean* específicos de las especies (al menos no debería sorprendernos descubrir que ése fuera el caso), y muy probablemente la conducta de nuestro marciano humanoide *esté* causada por sensaciones (o "schmensaciones" [*schmensations*]), diferentes, de alguna manera, de las nuestras, pese a las similitudes superficiales de su comportamiento con el nuestro.

No sé si alguien ha defendido explícitamente alguna vez la doctrina de los Dos Niveles [*Two-Levelism*] como tal.[21] Pero ella parece subyacer de modo directo a dilemas aparentes tales como "el problema de los *outputs* e *inputs*" que plantea Block.

Consideraciones paralelas valen para el problema de la intencionalidad. Pensamos que el estado de un organismo o es intencional o no lo es, y luego nos preguntamos cuál podría ser el lugar funcional o institucional de la intencionalidad. No creo que la intencionalidad pueda ser

20. "¡Esperemos hasta el año próximo!", se burlaba John Searle respecto de una conexión diferente, pero muy similar (1980). ¡*Por supuesto*, esperemos hasta el año próximo!

21. Salvo posiblemente el "funcionalista analítico", cuyo punto de vista rechazo (véase la nota 27).

una propiedad *puramente* funcional, por razones que ya son familiares,[22] pero en la medida en que lo pueda ser, creo que haríamos bien en admitir que la intencionalidad misma es gradual.[23] Las "señales" de la intencionalidad o de la referencia-a [*aboutness*] tampoco son muy claras, pero lo que parece evidente cuando reflexionamos es que existe un nivel intermedio de caracterización funcional que ofrece una *clase* de direccionalidad-hacia-un-objeto-o-tipo-no-existente-posible [*directedness-upon-a-possible-non-existent-object-or-type*] que cae sin embargo fuera de la intencionalidad plena y variada que exhibe la mente humana. En este nivel intermedio empleamos teóricamente términos sistémicos [*system-theoretical*], hablamos de "detectores", "lectores", "filtros", "inhibidores" y demás, entendiendo estos términos en sentido literal pero sin imputarles realmente *contenido* [*thought*] o lo que podría ser llamado referencia-a "ocurrente" [*"occurrent" aboutness*]. Pero tengo que dejar para otra ocasión el desarrollo de estas observaciones.[24]

En tercer lugar, podría ser beneficioso que nos atuviéramos a la caracterización "puramente abstracta" de *inputs* y *outputs*, remitiendo el problema del chauvinismo y el liberalismo a nuestra caracterización de los estados y eventos internos. Hay tantas posibilidades, tantos niveles de abstracción diferentes en la jerarquía funcional que se aplica al cerebro (muchos de los cuales se solapan y entrecruzan), que parece plenamente razonable esperar que haya, para cada estado mental tipo, alguna *vía* media entre el chauvinismo y el liberalismo, no necesariamente la *misma vía* para cada estado-tipo. Es simplemente un error pensar que todos los fenómenos mentales tengan que localizarse funcionalmente en el mismo nivel, o que un estado mental singular tenga que estar completamente localizado en un nivel. Atendiendo a los estados

22. Putnam (1975), Fodor (1980), Stich (1981), Burge (1979), Lycan (1981c),...

23. Dretske (1981) anticipa en parte esta idea. Véase también Van Gulick (1980; 1982).

24. Observaría también que algunas discusiones corrientes dentro de la comunidad de la ciencia cognitiva resultan mal planteadas dentro de la teoría de los Dos Niveles. Por ejemplo las discusiones entre los partidarios de la concepción "de-arriba-abajo" y los computacionalistas de la Iglesia Conservadora (véase P.S. Churchland, 1986 y Dennett, 1986), y entre estos últimos y los neo-conexionistas (véase Bechtel, 1985). Los neo-conexionistas en particular, constituyen un magnífico ejemplo de una *vía* media biocomputacional. El modelo de P.M. Churchland del "espacio de fases" de la coordinación senso-motora, basado en Pellionisz y Llinas (1979, 1982) está de alguna manera en este mismo espíritu; o más bien, aunque *el autor* no siempre piensa en un modo mediado, yo lo consideraría como otra vía media posible *dentro del espíritu de un funcionalismo propiamente teleologizado.*

mentales "más funcionales", *casi* netamente conductistas, tal vez no nos preocupe admitir que un sistema económico o la población de China podría tener tales estados (por ejemplo, creencias disposicionales), si fuera el caso. Y es posible que en el punto menos funcional del continuo haya incluso estados mentales tipo de los cuales la Teoría de la Identidad sea verdadera, aunque es difícil pensar que un estado mental sea tan poco "cualitativo" como ellos.

Las consideraciones anteriores sugieren una respuesta adicional al argumento del "*qualia* ausente" [*absent qualia*] de Block, respuesta que creo virtualmente concluyente. Antes caractericé la inquietud intuitiva de Block hacia el Funcionalismo como la cuestión de captar la incongruencia entre el carácter relacional de las explicaciones Funcionalistas y los rasgos cualitativos homogéneos primitivamente *monádico* de sus *explicanda*. Me imagino que esta incongruencia le parece absoluta. Nótese que, evidentemente, Block no tiene una objeción similar contra la Teoría de la Identidad; como cualquier otro materialista, simplemente le imputa chauvinismo, y nada más. Después de todo, una de las principales ventajas de la teoría era su habilidad para dar cuenta de la posibilidad del espectro invertido o de cualquiera otra variación interna, pese a la similitud exterior. Pero si aceptamos también mi tesis de que las caracterizaciones Homuncofuncionales y las caracterizaciones fisiológicas de los estados de las personas, meramente reflejan diferentes niveles de abstracción dentro de una jerarquía o continuo funcional abarcativo, entonces ya no podremos distinguir la Teoría de la Identidad de la Funcionalista de ninguna manera definitiva. "Neurona", por ejemplo, puede entenderse como un término fisiológico (que denota un tipo de célula humana) o como un término (teleo-)funcional (que denota un relé [*relayer*] de cargas eléctricas); en *cualquiera* de ambas versiones "neurona" designa algo instanciable —si se quiere—, un rol que es desempeñado por un grupo de objetos más fundamentales. En este sentido *incluso el teórico de la Identidad es Funcionalista*, un Funcionalismo que localiza las entidades mentales en un nivel de abstracción muy bajo. La moraleja es que si Block quiere realmente seguir sosteniendo que la psicología funcionalista está afectada por una incongruencia de principio del tipo ya mencionado, y que una filosofía de la mente que explique los ítemes mentales en términos de roles relacionales o instanciables no puede en principio acomodar el carácter irreduciblemente monádico de los *qualia*, entonces uno también debería formular el mismo cargo contra el teórico de la Identidad, y sospecho que Block no siente ninguna compulsión

intuitiva a hacerlo.[25] En realidad, deja que se acuse a la Teoría de un chauvinismo-de-especie de carácter global, e incluso permite que probablemente ése sea el caso respecto de algunas propiedades mentales.

Existe la idea, originada en un Bi-Nivelismo [*Two Levelism*] sin salida, de que el Funcionalismo difiere, *conceptual o estructuralmente*, de la Teoría de la Identidad, de un modo tal que está sujeto a clases diferentes de objeciones. Como dije, la Teoría de la Identidad es un caso empírico especial del Funcionalismo, un caso que (implausiblemente) localiza todos los estados mentales en un nivel extremadamente bajo de abstracción institucional: el neuroanatómico. De ese modo, no debería haber objeciones filosóficas o puramente conceptuales que se aplicaran al Funcionalismo y no a la Teoría de la Identidad, o viceversa, incluso si una fuera empíricamente menos razonable que la otra. Sin embargo filósofos como Block han pretendido ver tales objeciones. Si mi doctrina de la continuidad de la naturaleza es correcta, algo tiene que estar mal en ese punto; porque los términos neuroanatómicos son funcionales y tan relacionales como lo son los términos de las organizacionales superiores, aunque en un nivel de abstracción menor. Si existe una incongruencia de principio entre la caracterización relacional y el carácter intrínseco de las cualidades fenoménicas, y si esa incongruencia afecta al Funcionalismo, entonces debería afectar también a la Teoría de la Identidad.[26]

Consideremos un segundo ejemplo de tal objeción. Block sostiene además que el Funcionalismo es incapaz de admitir la posibilidad del "espectro invertido" u otros tipos de *qualia* cuyos cambios internos no se reflejen, siquiera contrafácticamente, en la conducta; incapaz en una manera en que no lo es la Teoría de la Identidad, dado que ella está *hecha a la medida* para representar casos de *qualia* invertidos [*inverted qualia*]. Pero si mis reflexiones sobre la continuidad de niveles de la naturaleza son correctas, algo errado tiene que haber en eso. Y algo hay. Así como es fácil imaginar cambios *neurofisiológicos* indetectables que subyacen al espectro invertido (véase Lycan, 1973), es fácil imaginar un cambio de componentes funcionales descriptos de manera más abstracta

25. Wilfrid Sellars lo hace. Pero ésa es otra historia...; véase el capítulo 8 de Lycan, 1987.

26. Block no acentúa el contraste monádico/relacional sino que ofrece sus intuiciones diferenciales puras; de ahí que pueda permanecer impasible ante mi anterior *ad hominem* e insistir simplemente en que tener una neuroquímica aproximadamente similar a la nuestra es condición necesaria para experimentar los *qualia*, relacionales o no. ¿Pero, cómo podría un *filósofo* saber eso?, me pregunto. ¿Brilla con Luz Natural?

(aunque sin duda hay límites a esto, y muy probablemente uno no podría ascender a un nivel muy elevado de abstracción y mantener la inversión indetectable en la conducta).

La verdad de la cuestión es oscurecida por una ambigüedad pragmática en la noción de "*qualia* invertido", una ambigüedad a la que Block ha dado ayuda retórica, aunque esté lejos de ser sutil. Hay un parámetro oculto, a saber: ¿"invertido" *respecto de qué*? (Compárese con la noción correlativa de *superveniencia*, ¿superveniente en qué?) Tradicionalmente el "espectro invertido" ha significado *qualia invertidos* (de color) sólo con respecto a relaciones *inputs-outputs* reales o contrafácticas. Sea por deber o por inclinación, los conductistas analíticos y los wittgensteinianos negaron la concebibilidad de *tal* inversión, pero las intuiciones modales ordinarias de la mayoría de la gente la favorecieron, y las teorías Funcionalista y de la Identidad también la acomodaron con facilidad. Ello jamás supuso un desafío al Funcionalismo. Lo que dañaría al Funcionalismo es la concebibilidad del *qualia* invertido con respecto a las relaciones I-O [*input-output*] *más* una organización funcional interna. *Esta* hipótesis acerca de la inversión es más fuerte y osada. Su posibilidad es al menos controvertible. En verdad, afirmarla es simplemente negar la verdad del Funcionalismo; es afirmar sin argumento que dos organismos podrían diferenciarse en sus estados cualitativos aun cuando fueran exactamente iguales en toda su organización funcional global, *cualquiera sea el nivel de abstracción institucional en cuestión*. Por supuesto que ha habido filósofos que insistieron sin argumentos en la posibilidad metafísica de organismos que difirieran en sus estados cualitativos pese a ser duplicados *moleculares*, pero tal insistencia carece de credibilidad intrínseca, aun si las teorías relevantes de la mente finalmente resultaran falsas. La posibilidad del espectro invertido sólo con respecto a relaciones I-O es una intuición modal bien asegurada y respetable, aunque supongo que refutable; la posibilidad del espectro invertido respecto de relaciones I-O *más una organización funcional interna en el nivel de abstracción que los proponentes crean plausible señalar, por bajo que sea*, no es obvia y está en conflicto con algunas tesis de superveniencia intuitivamente plausibles.

Algunos teóricos amantes de las relaciones pueden encontrar natural suponer un nivel de abstracción *privilegiado* desde el vamos. Por ejemplo, los "funcionalistas *analíticos*", o como prefiero llamarlos, los teóricos relacionales del sentido común, consideran que *los significados de los términos mentales* están determinados por sus roles causales asociados a esos términos por el sentido común o por la "psicología de sen-

tido común" ["*folk psychology*"], negándose con ello la posibilidad de apelar a cualquier nivel de organización funcional más bajo del que sea accesible al sentido común.[27] Dejando a un lado la psicología de sentido común, los computacionalistas de la "Iglesia Conservadora"[28] desdeñan apelar a la biología humana incluso en el marco de un enfoque puramente científico de la cognición y la conducta, aunque el nivel de la naturaleza que ellos eligen nunca es claramente especificado.[29] Un teórico que adhiera a un nivel de organización privilegiado puede por supuesto admitir el "espectro invertido" en relación con el nivel elegido, siempre que consienta en identificar —como digo— los *qualia* con ítemes de más bajo nivel aun.[30]

Dos estrategias alternativas

He recomendado un modo de resolver los problemas del chauvinismo y del liberalismo referentes a los *qualia* dentro de una ontología funcionalista de lo mental. Hay estrategias alternativas posibles. Un enfoque alternativo sería bifurcar nuestra concepción de lo mental, apropiándose, simplemente, de la distinción entre un estado mental y

27. En este punto estoy en deuda con Sydney Shoemaker, por su útil correspondencia. Por mi parte, no puedo aceptar el funcionalismo analítico, por dos razones: (i) rechazo la teoría del significado, sea del análisis conceptual o de la definición implícita, sobre la que esta teoría descansa. (Véase Armstrong, 1968 y Lewis, 1972 para las dos versiones más explícitas y sus correspondientes defensas, y Lycan, 1981b, especialmente la nota 10, para mi concepción alternativa de la semántica de los términos mentales; véase también Jacoby (1985), para un punto de vista similar). (ii) Dudo que el sentido común o la "psicología de sentido común" contenga información suficiente sobre las entidades mentales como para caracterizarlas de una manera tan rica como sería necesario para evitar contraejemplos. Sería muy fácil conseguir modelos armados [*clothespin*] de la psicología de sentido común, sin la masiva complejidad y la organización teleológica que la adscripción de estados mentales reales garantizaría.

28. El término se debe a Dennett (1986).

29. Sigo aquí la suposición de algunos escritores recientes de que existen realmente ciertos Computacionalistas de la Iglesia Conservadora; no estoy seguro de que ningún Funcionalista real haya mantenido concientemente esta posición. Se la adscribe usualmente a Zenon Pylyshyn y Jerry Fodor, sobre la base de algunas de sus observaciones sobre la realizabilidad múltiple. Quizá Ned Block la sostiene realmente, o de lo contrario no continuaría resistiéndose a mis argumentos en contra de la teoría de los Dos Niveles, como la planteé en mi "Form, function and feel", 1981a.

30. Véase en particular (nuevamente) Bechtel (1985), y las referencias que figuran allí.

su carácter cualitativo, explicando los estados en términos funcionales y los caracteres en términos psicológicos muy amplios, tolerando la consecuencia de que el problema del espectro invertido o ciertas diferencias interpersonales menores en los *qualia* pudieran ser más comunes de lo que pensamos (por ejemplo, tan comunes como lo son las diferencias fisiológicas interpersonales de magnitud comparable).[31]

El *dolor* podría constituir un caso testigo para esta segunda manera de tratar de acomodar a los *qualia*. Algo interesante y distintivo del dolor es que (a diferencia de la mayoría de los estados mentales) tiene a la vez, un fuerte esquema de conducta asociado, y vívidas propiedades experimentables introspectibles. Esto significa, en tal propuesta, que los estados de dolor pueden recibir análisis en *muchos niveles* [*multileveled*]. Podríamos, por ejemplo (especulando un poco), terminar clasificando como dolor cualquier estado interno de un organismo que ocupara el rol conductual "grueso" ["*gross*"] (que el dolor usualmente tiene): el de ser causado por algo dañoso y de producir una retracción-con-preferencia [*withdrawal-cum-favoring*], pero distinguir las sensaciones de dolor según las bases fisiológicas de los estados internos.[32] De ello seguiría que aunque los moluscos y los marcianos sufran dolor, lo sienten probablemente de manera distinta de como nosotros sentimos el nuestro. También seguiría que un estado que produzca la sensación que produce en mí el dolor, podría ser un estado mental diferente del de dolor en una criatura con una organización distinta; algunos filósofos pueden hallar esto crasamente contraintuitivo.

Incidentalmente, la concepción bifurcada se ha vuelto bastante popular en los años recientes,[33] y se la expresa a menudo diciendo que (por ejemplo) "el dolor en sí mismo es funcional mientras que su sensación específica es neurofisiológica". Pero esta formulación nuevamente presupone la doctrina de los dos niveles. Considérese a la distinción "funcional"/"estructural" como relativa-a-nivel, y la teoría bifurcada colapsa en una versión específica poco aguda de la tesis (que espero se torne una verdad obvia) de que los estados mentales y sus características

31. Block (1978) insinúa que no congenia del todo con esta sugerencia. Véase también nota 33.

32. Esta movida quitaría algo de peso a lo que entiendo que es un argumento anti-Funcionalista en David Lewis, 1980.

33. Como ya mencioné, Block parece inclinarse hacia esta posición. Desarrollo esta sugerencia en Lycan, 1981a, págs. 47-8. Esta idea ha sido recogida también por Hilary Putnam (1981), Sidney Shoemaker (1981), Patricia Kitcher (1982), Terence Horgan (1984) y Gregory Sheridan (1986), entre otros (Shoemaker la denomina "estrechez selectiva").

cualitativas no pueden ser explicados en términos del mismo nivel de la naturaleza (en particular, la ubicación del carácter cualitativo puede ser más baja en la jerarquía que la del estado mental genéricamente considerado). Estoy totalmente de acuerdo con esta tesis, como ya lo indiqué, pero ella no es una *alternativa* que compita con el Funcionalismo.

Un tercer enfoque alternativo para el caso de las sensaciones corporales se sugiere por sí mismo (aunque dudo que pueda ser aplicado fácilmente a los *qualia* perceptuales). Se supone que las sensaciones que parecen fenoménicamente simples, son realmente complejas y que el *quale* distintivo asociado a una sensación de un cierto tipo, es realmente la coincidencia o superposición de distintos rasgos homunculares individualmente manejables. Pienso que esta línea, más bien que la bosquejada en el párrafo anterior, es la más plausible de adoptar para el caso del dolor, porque es fuertemente sugerida por los datos anestesiológicos reunidos y resumidos por Dennett (1978, cap. 11). Lo que estos datos parecen indicar es que anestésicos y analgésicos químicamente diferentes interrumpen las subrutinas normales de "dolor" de los sujetos en diferentes etapas funcionales, provocando informes verbales totalmente distintos de los efectos. En un grupo de individuos que sufren dolor aproximadamente de la misma intensidad, un subgrupo al que se le da la droga A puede informar que el dolor ha disminuido o ha cesado por completo, mientras que otro subgrupo al que se le dio la droga B puede informar que aunque sabe que el dolor aún existe, no lo puede sentir; y los miembros de un subgrupo a quienes se dio la droga C pueden informar que aunque todavía pueden sentir el dolor tan intensamente como antes, no les *importa*, y así sucesivamente. Que algunos de esos informes nos suenen cómicos (podrían ser rechazados como ininteligibles por algunos wittgeinsteinianos) refleja naturalmente el hecho de que se han interrumpido los procesos internos normales de los miembros del grupo, y que sus experiencias internas normales de dolor han sido alteradas por las drogas. Lo que las drogas parecen hacer es *separar los componentes* de la experiencia fenoménica que los sujetos tienen del dolor, al separar las sub-subrutinas de sus bases funcionales más complicadas. Y si esto es así, se sigue que nuestra experiencia fenoménica de dolor tiene componentes; es un complejo que consiste (tal vez) en pulsiones instintivas, deseos, impulsos y creencias, que probablemente acaezcan en niveles de abstracción funcional totalmente diferentes. Si estos componentes pueden ser separados individualmente unos de otros por medio de drogas, entonces podemos realizar un *Gedankenexperiment* en el cual tomamos hipotéticamente a un sujeto que sufre, sepa-

ramos un componente de su dolor administrándole la droga A, luego separamos otro componente administrándole la droga B, y repetimos el proceso generando informes, para no perder de vista el modo en que lo estamos haciendo. Me parece plausible pensar que si siguiéramos este procedimiento, interrumpiendo una vía de acceso tras otra y eliminando los componentes de pulsiones instintivas, deseos y creencias uno por uno, lograríamos tarde o temprano eliminar el dolor mismo; parece también que si revirtiéramos el proceso —restaurando las vías de acceso al suspender las diversas drogas una a una— necesariamente el sujeto sentiría de nuevo la sensación de dolor primitiva (suponiendo que sus tejidos dañados no hubieran sanado en el ínterin). Creo que esto hace razonable suponer (nuevamente) que para que el dolor acaezca es necesaria y suficiente —en contra del espíritu antiliberal de Block— alguna subsecuencia apropiada de *varios niveles* del complejo relevante de conductas funcionales.

No sé cómo elegir concluyentemente entre los tres enfoques alternativos que describí, o qué clase de evidencia ulterior podríamos buscar. Recorrí algunas de las opciones sólo para mostrar que el Homuncofuncionalista tiene una gran riqueza de recursos para lidiar tanto con el dilema del chauvinismo y el liberalismo como con la tarea positiva de dar cuenta de los *qualia*. Sobre la base de estos recursos creo que podemos concluir que el pesimismo de Block sobre los *qualia* es incorrecto...

Si mi doctrina de la continuidad, tal como la formulé, es obvia, no ha sido lo suficientemente obvia para algunos de nuestros principales filósofos de la mente. Espero que ulteriores demostraciones sirvan para hacer más atractivo, como teoría de la mente, al Homuncofuncionalismo.

TRADUCTORES: Alejandro Miroli y Alicia Pazos.
REVISIÓN TÉCNICA: Eduardo Rabossi.

REFERENCIAS BIBLIOGRÁFICAS

Armstrong, D.M.: (1968) *A Materialist Theory of the Mind*, Routledge & Kegan Paul.

Attneave, F.: (1960) "In defense of homunculi", en W. Rosenblith (comp.), *Sensory Communication*, MIT Press.

Bechtel, P.W.: (1985) "Contemporary connectionism: Are the new para-

llel distributed processing models cognitive or associationist?", *Behaviorism* 13, 53-61.

Bennet, J.: (1976) *Linguistic Behaviour*, Cambridge University Press.

Bigelow, J. y Pargetter, R.: (1987) "Functions", *Journal of Philosophy* 84, 181-96.

Biro, J.: (1981) "Persons as corporate entities and corporation as persons", *Nature and System* 3, 173-80.

Block, N.J.: (1978) "Troubles with functionalism", en W. Savage (comp.) *Perception and Cognition: Minnesota Studies in the Philosophy of Science*, Vol. 9, University of Minnesota Press.

Block, N.J.: (1981) "Psychologism and behaviorism", *Philosophical Review* 90, 5-43.

Brooks, D.H.M.: (1986) "Group minds", *Australasian Journal of Philosophy* 64, 456-70.

Burge, T.: (1979) "Individualism and the mental", en P. French, T.E. Uehling y H. Wettstein (comps.), *Midwest Studies in Philosophy, Vol. IV: Studies in Metaphysics*, University of Minnesota Press.

Churchland, P.M.: (1986) "Some reductive strategies in cognitive neurobiology", *Mind* 95, 223-38.

Churchland, P.S.: (1986) *Neurophilosophy*, Bradford Books/MIT Press.

Cummins, R.: (1983) *The Nature of Psychological Explanation*, Bradford Books/MIT Press.

Davidson, D.: (1970) "Mental events", en L. Foster y J.W. Swanson (comps.), *Experience and Theory*, University of Massachusetts Press.

Dennett, D.C.: (1975) "Why the law of effect will not go away", *Journal of the Theory of Social Behaviour*, págs. 169-176.

Dennett, D.C.: (1978) *Brainstorms*, Bradford Books.

Dennett, D.C.: (1986) "The logical geography of computational approaches: a view from the East Pole", en M. Brand y R.M. Harnish (comps.), *The Representation of Knowledge and Belief*, University of Arizona Press.

Dretske, F.: (1981) *Knowledge and the Flow of Information*, Bradford Books/MIT Press.

Fodor, J.A.: (1968) "The appeal to tacit knowledge in psychological explanation", *Journal of Philosophy*, 65, 627-40.

Fodor, J.A.: (1980) "Methodological solipsism considered as a research strategy in cognitive psychology", *Behavioral and Brain Sciences* 3, 67-73.

French, P.: (1984) *Collective and Corporate Responsibility*, Columbia University Press.

Horgan, T.: (1964) "Functionalism, *qualia*, and the inverted spectrum", *Philosophy and Phenomenological Research* 44, 453-70.

Jacoby, H.: (1985) "Eliminativism, meaning and qualitative states", *Philosophical Studies* 47, 257-70.

Kitcher, P.: (1982) "Two versions of the Identity Theory", *Erkenntnis* 17, 213-28.

Lewis, D.: (1972) "Psychophysical and theoretical identifications", *Australasian Journal of Philosophy* 50, 249-58.

Lewis, D.: (1980) "Mad pain and Martian pain", en N. Block (comp.), *Readings in Philosophy of Psychology,* Vol. 1, Harvard University Press.

Lycan, W.: (1973) "Inverted spectrum", *Ratio* 15, 315-19.

Lycan, W.: (1981a) "Form, function and feel", *Journal of Philosophy* 78, 24-50.

Lycan, W.: (1981b) "Psychological laws", *Philosophical Topics* 12, 9-38.

Lycan, W.: (1981c) "Toward a homuncular theory of believing", *Cognition and Brain Theory* 4, 139-59.

Lycan, W.: (1984) *Logical Form in Natural Language,* Bradford Books/MIT Press.

Lycan, W.: (1987) *Consciousness*, Bradford Books/MIT Press.

Lycan, W.: (1989) "Ideas of representation", en *Mind, Value and Culture: An Essay in Honor of E. M. Adams* (compilado por D. Weissbord), Ridgeview Publishing.

MacKenzie, A.W.: (1972) "An analysis of purposive behavior", Tesis doctoral, Cornell University.

Matthen, M. y Levy, E.: (1984) "Teleology, error, and the human immune system", *Journal of Philosophy* 81, 351-71.

Mellick, D.: (1973) "Behavioral strata", Tesis doctoral, Ohio State University.

Millikan, R.G.: (1984) *Language, Thought, and Other Biological Categories*, Bradford Books/MIT Press.

Neander, K.: (1981) "Teleology in biology", Fotocopia inédita.

Neander, K.: (1983) "Abnormal psychobiology", Tesis doctoral, La Trobe University.

Pellionisz, A. y Llinas, R.: (1979) "Brain modelling by tensor network theory and computer simulation. The cerebullum: distributed processor for predictive coordination", *Neuroscience* 4, 323-48.

Pellionisz, A. y Llinas, R.: (1982) "Space-time representation in the

brain. The cerebellum as a predictive space-time metric tensor", *Neuroscience* 7, 2949-70.

Popper, K.: (1972) "Of clouds and clocks: An approach to the problem of rationality and the freedom of Man", en *Objective Knowledge: An Evolutionary Approach*, Oxford University Press.

Putnam, H.: (1975) "The meaning of 'meaning'", en K. Gunderson (comp.), *Minnesota Studies in the Philosophy of Science, Vol. 7: Language, Mind and Knowledge*, University of Minnesota Press.

Putnam, H.: (1981) *Reason, Truth and History*, Cambridge University Press.

Richardson, R.: (1983) "Computational models of mind." Monografía inédita.

Searle, J.: (1980) "Minds, brains and programs", *Behavioral and Brain Science* 3, 417-24.

Sheridan, G.: (1983) "Can there be moral subjects in a physicalistic universe?", *Philosophy and Phenomenological Research* 43, 425-48.

Sheridan, G.: (1986) "Selective Parochialism and Shoemaker's Argument for Functionalism", manuscrito. Western Michigan University.

Shoemaker, S.: (1981) "Some varieties of Functionalism", *Philosophical Topics* 12, 93-119.

Simon, H.: (1969) "The Architecture of complexity", en *The Science of the Artificial*, MIT Press.

Sober, E.: (1985) "Panglossian Functionalism and the Philosophy of Mind", *Synthèse* 64, 165-93.

Stich, S.: (1981) "Dennett on intentional systems", *Philosophical Topics*, 12, 39-62.

Thomas, L.: (1974) *Lives of a Cell*, Bantam Books.

Van Gulick, R.: (1980) "Functionalism, information and content", *Nature and Systems* 2, 139-62.

Van Gulick, R.: (1982) "Mental representation —a functionalist view", *Pacific Philosophical Quarterly* 63, 3-20.

Wimsatt, W.C.: (1972) "Teleology and the logical structure of function statements", *Studies in History and Philosophy of Science* 3, 1-80.

Wimsatt, W.C.: (1976) "Reductionism, levels of organization, and the mind-body problem", en G. Globus, G. Maxwell e I. Savodnik (comps.), *Consciousness and the Brain*, Plenum.

Wright, L.: (1973) "Functions", *Philosophical Review* 82, 139-68.

IV

LAS ACTITUDES PROPOSICIONALES Y EL LENGUAJE DEL PENSAMIENTO

CAPÍTULO 6

LAS ACTITUDES PROPOSICIONALES *

Jerry A. Fodor

Algunos filósofos (John Dewey, por ejemplo, y tal vez John Austin) sostienen que la filosofía es lo que uno hace con un problema hasta que se torna lo suficientemente claro como para que la ciencia lo resuelva. Otros filósofos (Gilbert Ryle, por ejemplo, y tal vez Ludwig Wittgenstein) sostienen que si un problema filosófico sucumbe a los métodos empíricos, ello muestra que desde un comienzo no era *realmente* filosófico. De cualquier manera, los hechos son bastante claros: problemas inicialmente tratados por los filósofos suelen caer en manos de personas que hacen experimentos. Esto parece estar ocurriendo hoy en día con la pregunta '¿Qué son las actitudes proposicionales?', y la ciencia prestigiosa es la psicología cognitiva.

Una manera de elucidar esta situación es examinando las teorías que los psicólogos cognitivistas aceptan, con miras a explicar la descripción de las actitudes proposicionales que esas teorías presuponen. Esa fue mi estrategia en Fodor (1975). En este trabajo, en cambio, tomaré otro rumbo. Quiero bosquejar ciertas condiciones a priori que, en mi opinión, deben ser satisfechas por una teoría de las actitudes proposicionales (APs). Argumentaré que, en conjunto, esas condiciones exigen claramente tratar a las APs como relaciones entre organismos y representaciones internas; ésta es, precisamente, la concepción a la que los psicólogos han llegado de manera independiente. Argumentaré entonces que, aun dejando a un lado las exigencias empíricas que los llevaron a ella, tenemos buenas razones para aceptar la teoría de los psicólogos. Considero que la convergencia entre lo que es plausible a priori y lo que es exigido *ex post facto* es en sí misma una razón para creer que probablemente la teoría sea verdadera.

Tres comentarios preliminares. Primero, no estoy tomando en serio la expresión "a priori". Algunos de los puntos que defenderé son,

* "Propositional Attitudes", *The Monist* 61 (1978), págs. 501-523. Con autorización del autor y de *The Monist*.

supongo, estrictamente conceptuales, pero otros son meramente autoevidentes. Lo que he obtenido es un conjunto de hechos notorios acerca de las actitudes proposicionales. Pero no me caben dudas de que podríamos aceptar racionalmente una explicación de las actitudes que contradiga algunos o incluso todos esos hechos. Pero la evidencia independiente a favor de tal explicación tendrá que ser extremadamente persuasiva, o yo, al menos, no estaré tranquilo. Segundo, prácticamente todo lo que diga acerca de las actitudes ha sido dicho previamente en la literatura filosófica. Todo lo que he hecho es juntar las cosas. Pienso, sin embargo, que las diferentes condiciones que discutiré se iluminan mutuamente: sólo cuando se intenta satisfacerlas a todas al mismo tiempo, uno advierte cuán unívocas son sus exigencias. Finalmente, si bien considero que lo que digo se aplica, mutatis mutandis, al conjunto total de las APs, centraré la discusión casi exclusivamente en las creencias y los deseos. Ellos parecen ser los casos básicos para una psicología cognitiva sistemática; el aprendizaje y la percepción deben pues ser tratados, presumiblemente, como variedades de la fijación de las creencias, y la teoría de la acción probablemente sea contigua con la teoría de la utilidad.[1]

Aquí van pues mis condiciones, con comentarios.

I. Las actitudes proposicionales deben ser analizadas como relaciones. En particular, el verbo en una oración como 'Juan cree que está lloviendo' expresa una relación entre Juan y algo, y un ejemplar [*token*] de esa oración es verdadero, si y sólo si Juan está en la relación-de-creencia con ese algo.[2] Consiguientemente, y a esos efectos, en 'Juan cree que está lloviendo', 'está lloviendo' funciona como término.[3] Tengo tres argumentos para imponer la condición I, ninguno de los cuales es concluyente.

1. No diré nada acerca de conocer, descubrir, reconocer, o cualquiera de las actitudes "fácticas" ["*factive*"]. La justificación de esta restricción merece ser discutida, pero no aquí.

2. No distingo la tesis relacional de la tesis que considera a 'Juan cree' como un operador que actúa sobre 'está lloviendo'; la parte que más me interesa es 'está lloviendo', no 'Juan cree'.

3. Supongo que lo que sigue es aproximadamente correcto: dada una oración cuya forma sintáctica es $NP_1\ (V(NP_2))_{VP}$, V expresa una relación si y sólo si NP_1 y NP_2 refieren. Por lo tanto, para nuestros fines, el problema de si 'cree' expresa una relación en 'Juan cree que está lloviendo' equivale al problema de si existen o no cosas tales como los objetos de las creencias. Por consiguiente, en la discusión que sigue, no me molestaré en distinguir entre las múltiples maneras posibles de expresar el problema.

Ia. Es intuitivamente plausible: 'cree' parece ser una relación binaria, y sería bueno que nuestra teoría de la creencia nos permitiera salvar las apariencias.

Sin duda, las apariencias a veces engañan. El 'de' en '[por] el bien de María' [*Mary's sake*] parece expresar una relación (de posesión) entre María y el bien; pero no es así, o, por lo menos, eso es lo que se dice. De hecho, 'el bien de María' no parece *muy* relacional, dado que *el bien de x* sería considerado sin duda como un giro idiomático aun cuando no tuviéramos ningún escrúpulo ontológico que aplacar. Hay algo sintácticamente erróneo en: 'el bien de María es mejor que el de Guillermo', 'María tiene un bien menor', etcétera. Lo cierto es que 'un bien' conlleva un error sintáctico, *tout court*. Sin embargo, esperaríamos que todas esas expresiones estuvieran bien formadas si 'el bien de María' incluyera un posesivo verdadero. 'El bien de María' no puede ser comparado con 'el cordero de María'.

Sin embargo, hay casos de expresiones *no* idiomáticas que parecen ser relacionales, pero que, al cabo de cierta reflexión, tal vez no lo son. 'la voz de María' admite las transformaciones descartadas en el caso de 'el bien de María' (Dennett, 1969). Sin embargo no hay, tal vez, *cosas* tales como voces; y, si no las hay, 'la voz de María' no puede referir en virtud de la relación entre María y alguna de tales cosas.[4] Pienso que, en *estos* casos, es adecuado considerar a la gramática 'superficial' como ontológicamente engañosa, pero sólo porque sabemos cómo traducirlos a formas más parsimoniosas. 'María tiene una buena voz (mala voz; poca voz; mejor voz que Guillermo)' se puede transformar sin pérdida de significado en 'María canta bien (mal; sin fuerza; peor que Guillermo)'. Pero si *no* estuviéramos en condiciones de proveer las traducciones adecuadas (o, aun, de visualizar cómo proveerlas), ¿qué derecho tendríamos a considerar dichas expresiones como ontológicamente promiscuas? 'Guillermo cree que está lloviendo' no es un giro idiomático, y no hay, por lo que se sabe, manera alguna de traducir oraciones acerca de creencias a oraciones con una carga ontológica reducida. (Los conductistas solían pensar que dichas traducciones eran posibles, pero estaban equivocados.) Por lo tanto, tenemos que tomar en serio los compro-

4. Por supuesto, podría referir en virtud de una relación entre María y algo distinto de una voz. 'Juan es más alto que el hombre medio' no es verdadera en virtud de una relación entre Juan y el hombre medio ('el hombre medio' no refiere). Pero la oración es igualmente relacional. Es por este tipo de razones que principios tales como el enunciado en la nota 3 son sólo aproximadamente válidos.

misos ontológicos aparentes o admitir que somos intencionadamente descuidados.

Ib. La Generalización Existencial (GE) se aplica a los objetos sintácticos de los verbos de actitud proposicional; de 'Juan cree que está lloviendo' es posible inferir 'Juan cree algo' y 'Hay algo que Juan cree' (esto es, que está lloviendo). La GE podrá no ser un *criterio* de compromiso ontológico, pero es ciertamente un indicio.[5]

Ic. La única alternativa conocida a la tesis de que los verbos de actitud proposicional expresan relaciones, es que están (semánticamente) "fusionados" [*fused*] con sus objetos, y una concepción tal no parece tener futuro.[6]

La teoría consiste en proponer que oraciones como 'Juan cree que está lloviendo' deberían realmente escribirse 'Juan cree-que-está-lloviendo'; que la forma lógica de esas oraciones contiene una expresión referencial ('Juan') y un predicado monádico sin estructura interna ('cree-que-está-lloviendo'). 'Juan cree que está lloviendo' es una oración atómica, similar, en definitiva, a 'Juan está morado'.

¡Y hablan de cosas antiintuitivas! Además:

1. Hay un número infinito de oraciones (semánticamente distintas) de la forma *a cree complemento*. Si todas ellas son atómicas, ¿cómo se puede aprender el castellano? (Davidson, 1965).

5. N.B.: los verbos de actitud proposicional son transparentes en este sentido sólo cuando sus objetos son *complementos*; no es posible inferir 'Hay algo que Ponce de León buscó' a partir de 'Ponce de León buscó la Fuente de la Juventud'. Sin embargo, tal vez valga la pena traducir 'buscar' por 'tratar de encontrar', y salvar de este modo la generalización. Esto nos daría: 'Ponce de León trató de encontrar la Fuente de la Juventud', lo cual supongo que *sí* implica que hay algo que Ponce de León trató (es decir, trató de hacer; a saber, encontrar la Fuente de la Juventud). Además, decir que la GE se aplica *al* complemento de los verbos de AP_s no es, por supuesto, decir que se aplica *en* el complemento de los verbos de AP_s. 'Juan quiere casarse con María de Rumania' implica que hay algo que Juan quiere (es decir, quiere hacer, a saber, casarse con María de Rumania); pero obviamente *no* implica que hay alguien con quien Juan quiere casarse (véase III abajo).

6. En varios contextos filosóficos, la fusión ha sido considerada como una solución *para* la falta de *transparencia*; véase Nagel, 1965; Goodman; Dennett, 1969. Nótese: 'considerada', no 'adoptada'.

2. Distintas actitudes proposicionales están centradas, a menudo, en el mismo contenido; por ejemplo, uno puede al mismo tiempo temer y desear que el martes llueva. Pero, según la teoría de la fusión, 'Juan teme que el martes llueva' no tiene nada en común con 'Juan desea que el martes llueva', exceptuando la referencia a Juan. En particular, es un *accidente* que la expresión lingüística 'el martes llueva' aparezca en ambas oraciones.

3. Además, creencias distintas pueden relacionarse de la siguiente manera: Juan piensa que Samuel es simpático; María piensa que Samuel es desagradable. Según la manera corriente de representar en castellano, esas creencias se superponen en la posición de 'Samuel', por lo que la notación sustenta la intuición de que Juan y María están en desacuerdo acerca de Samuel. Pero, si la tesis de la fusión es correcta, 'Juan piensa que Samuel es simpático' y 'María piensa que Samuel es desagradable' tienen tanto en común en el nivel de la notación canónica como, por ejemplo, 'Juan come' y 'María nada'. ¡Y hablan de cosas oscuras! En cuanto a sustentar intuiciones, la reconstrucción recomendada es *peor* que la descuidada ortografía con la que comenzamos.[7] (Lo cierto es que no hay nada en cree-que-S que sugiera que se refiere a una creencia. También en este sentido cree que S es preferible.)

4. Es difícil que sólo sea un accidente que las oraciones declarativas del castellano sean los objetos (sintácticos) de verbos como 'creer'. Pero, según la tesis de la fusión, ése es *precisamente* el caso; el complemento de 'creer' en 'Juan cree que está lloviendo' tiene la misma relación con la oración 'Está lloviendo' que, digamos, la palabra 'lago' con las dos últimas sílabas de 'murciélago'.

5. De acuerdo con la tesis de la fusión, es un mero accidente que si 'Juan cree que está lloviendo' es verdadera, entonces lo que Juan cree es verdadero si y sólo si 'Está lloviendo' es verdadera. Pero esto implica admitir demasiados accidentes. Sin duda, la identidad entre las condiciones de verdad de la creencia de Juan, cuando cree *Fa*, y las condiciones de verdad de la oración correspondiente *a* es *F*, tiene que ser lo que conecta la teoría de la interpretación de las oraciones con la teoría de

7. 3 no es acerca de la GE. En la teoría de la fusión, el hecho de que 'la creencia de que Samuel es simpático' es acerca de Samuel, no es representado, aun cuando 'creencia' y 'acerca de' sean ambos interpretados *de manera opaca*.

las APs (y lo que explica que usemos 'Está lloviendo', en lugar de cualquiera otra oración, para especificar *qué* creencia tiene Juan cuando cree que está lloviendo).

Hacer que los datos parezcan fortuitos es característico de las malas teorías. Concluyo, pues, que la tesis de la fusión no debe ser tomada muy en serio, y que ni la filosofía del lenguaje ni la filosofía de la mente pueden avanzar haciendo proliferar guiones. Pero la tesis de la fusión es (*de hecho*) la única alternativa a la tesis de que 'creer' expresa una relación. Por lo tanto, en principio, nos convendría suponer que 'creer' expresa una relación. Consecuentemente, en principio, nos convendría suponer que 'creer' expresa, *efectivamente*, una relación, y tratar de ofrecer una explicación de las actitudes proposicionales que concuerde con esta suposición.

II. Una teoría de las APs debería explicar el paralelismo existente entre los verbos de APs y los verbos que corresponden a decir ("la condición de Vendler").

En general, las cosas que podemos decir que *creemos* (deseamos, confiamos, lamentamos, etcétera) son las mismas cosas de las que podemos decir que *decimos* (aseveramos, enunciamos, etcétera). Así, Juan puede tanto creer como aseverar que está por soplar un vendaval; puede tanto confiar en que alguien esté tomando los rizos de la vela mayor como preguntar si alguien lo hace; puede tanto dudar de que los tripulantes abandonen el Genny como exigir que lo hagan. Además, tal como ha mostrado Zeno Vendler (1972), se obtienen consecuencias interesantes al clasificar los verbos de APs (de un lado) y los verbos de decir (del otro), tomando en cuenta la sintaxis de sus complementos de objeto. Las taxonomías así generadas resultan isomórficas hasta en niveles de análisis sorprendentemente refinados. Por supuesto, esto *podría* ser nada más que un accidente, como los paralelismos semánticos y sintácticos que se dan entre los complementos de los verbos de APs y las oraciones declarativas no subordinadas. Por cierto que Vendler formula una inferencia substancial al pasar de las similitudes sintácticas observadas a la conclusión que extrae, a saber, que el objeto de la aserción es idéntico al objeto de la creencia. Bástenos por el momento formular una propuesta menos ambiciosa: deberíamos preferir una teoría que explique los hechos a una que meramente se encoja de hombros; con otras palabras, deberíamos preferir una teoría que satisfaga la condición de Vendler a una que no lo haga.

III. Una teoría de las actitudes proposicionales debería dar cuenta de la opacidad ("la condición de Frege").

Hasta el momento, he enfatizado las analogías lógico-sintácticas entre los complementos de las cláusulas de creencia y las correspondientes oraciones declarativas no subordinadas. Sin embargo, desde los tiempos de Gottlob Frege, en la literatura filosófica se acostumbra enfatizar una de sus *diferencias* más notables: las primeras son, por lo general, opacas en las operaciones inferenciales en las que las segundas son, generalmente, transparentes. Dado que este aspecto del comportamiento de las oraciones que adscriben actitudes proposicionales ha dominado la discusión filosófica, aquí seré breve. Las oraciones que contienen verbos de APs no son, normalmente, funciones de verdad de sus complementos. Más aun, los contextos subordinados a los verbos de APs no son normalmente veritativo-funcionales, y la GE y la sustitución de idénticos pueden valer en las posiciones sintácticas de una oración declarativa no subordinada, mientras que no valen en posiciones sintácticamente comparables de las oraciones de creencia. Una teoría de las APs debería explicar por qué esto es así.

Debe reconocerse que, a pesar de sus groseras insuficiencias, la tesis de la fusión provee al menos una descripción de las actitudes proposicionales que cumple con la condición de Frege. Si *S* no aparece en #Juan cree *S*# no es de sorprender que una no sea una función de verdad de la otra; de la misma manera, si 'María' no aparece en 'Guillermo cree que Juan engañó a María', no es de sorprender que la oración no se comporte del modo en que se comportaría si 'María' ocurriera de manera referencial. La moraleja metodológica es, quizá, que la condición de Frege subrestringe [*underconstrains*] a una teoría de las APs; lo ideal es que una explicación adecuada de la opacidad siga de una teoría que es plausible por razones independientes.

IV. Los objetos de las actitudes proposicionales tienen forma lógica ("la condición de Aristóteles").

Los estados mentales (incluso, especialmente, los casos de actitudes proposicionales) interactúan causalmente. Esas interacciones constituyen los procesos mentales que resultan (*inter alia*) en las conductas de los organismos. Ahora bien, para un programa que se propone explicar la conducta en términos de estados mentales, es crucial que las actitudes proposicionales pertenecientes a esas cadenas estén típicamente relacionadas, en lo que respecta a su contenido, de manera *no* arbitraria (considerando, informalmente, que el contenido de una actitud proposicio-

nal esté constituido por lo que expresa el complemento de la correspondiente oración adscriptora de APs).

Esta no es una afirmación a priori, aunque tal vez sea trascendental. Porque, si bien uno puede imaginar la ocurrencia de cadenas causales de estados mentales relacionados de otras maneras (como, por ejemplo, el pensamiento de que dos es un número primo causa un deseo de beber té, que causa la intención de recitar el alfabeto de atrás para adelante, que causa la expectativa de lluvia) y si bien estas secuencias ocurren de hecho (en el sueño, por ejemplo, y en la locura), sin embargo, si toda nuestra vida mental fuera así, es difícil ver qué sentido tendrían las adscripciones de contenido a los estados mentales. Aun la fenomenología presupone alguna correspondencia entre el contenido de nuestras creencias y el contenido de nuestras creencias acerca de nuestras creencias; de lo contrario, los fenomenólogos no podrían dar cuenta de ninguna introspección coherente.

La situación paradigmática —el grano para el molino cognitivista— es aquella en la cual las actitudes proposicionales interactúan causalmente *en virtud de* su contenido. Y el paradigma de este paradigma es el silogismo práctico. Dado que mi tesis es, en parte, que los detalles no importan, me tomaré ciertas libertades con el texto aristotélico.

> Juan cree que va a llover si lava su auto. Juan quiere que llueva. Por lo tanto, Juan intenta que su acción sea un lavado de auto [*a car-washing*].

Considero que ésta podría ser una etiología verdadera, aunque informal, de la conducta de Juan; el lavar el auto es efecto de la intención de lavar el auto, y la intención de lavar el auto es efecto de la interacción causal entre las creencias de Juan y sus necesidades (conveniencias). Más aún, la explicación etiológica podría tener apoyo contrafáctico, al menos en el siguiente sentido: Juan no habría lavado el auto si el contenido de sus creencias, necesidades e intenciones hubiera sido distinto del que fue. Y si lo hubiera lavado, no lo habría hecho intencionadamente, o lo habría hecho por otras razones, o con otros fines. Decir que los estados mentales de Juan interactúan causalmente *en virtud de* su contenido es, en parte, afirmar la validez de tales contrafácticos.

Si hay contrafácticos verdaderos y contingentes que relacionan a los *casos* de los estados mentales en virtud de sus contenidos, es porque hay, presumiblemente, generalizaciones verdaderas y contingentes que relacionan estados mentales *tipo* [*types*] en virtud de sus contenidos. Por lo

tanto, siguiendo aun a Aristóteles a cierta distancia, es posible esquematizar etiologías como la anterior, de manera de obtener las generalizaciones subyacentes: si *x* cree que *A* es una acción que *x* puede realizar; y si *x* cree que la realización de *A* es suficiente para causar que *Q*; y si *x* quiere que sea el caso que *Q*; entonces, *x* actúa de una manera que tiene como propósito realizar *A*.

No estoy interesado en investigar aquí si ésta es una teoría plausible de la decisión; menos aún si es la teoría de la decisión que Aristóteles consideraba plausible. Lo que me interesa señalar es más bien: (a) que cualquier teoría de la decisión que consideremos se parecerá a ésta en que (b) implicará generalizaciones acerca de las relaciones causales entre creencias, necesidades e intenciones relacionadas por el contenido y (c) que esas generalizaciones serán especificadas con referencia a la forma de las actitudes proposicionales que las instancian. (Esto sigue siendo verdadero aun cuando, como algunos filósofos suponen, una adecuada teoría de la decisión requiere necesariamente cláusulas *ceteris paribus* para dar substancia a sus generalizaciones. Véase, por ejemplo, Grice, 1975.) Por lo tanto, en particular, no podemos enunciar la generalización teórica relevante que es instanciada por las relaciones entre los estados mentales de Juan, a menos que admitamos hacer referencia a creencias de la forma *si X entonces Y*; a deseos de la forma *que Y*; a intenciones de la forma *que X ocurra*, etcétera. Vistas de cierta manera (según el modo material), las letras esquemáticas recurrentes exigen identidades de contenido entre las actitudes proposicionales. Vistas de otra manera (lingüísticamente), requieren identidades formales entre los complementos de la oración adscriptora de APs que instancia las generalizaciones de la teoría que explica la conducta de Juan. Desde ambos puntos de vista, la forma de la generalización determina cómo se relaciona la teoría con los eventos que subsume. No hay nada notable en ello, por supuesto, salvo que la forma se adscribe *dentro* del alcance de los verbos de APs.

En síntesis: las generalizaciones psicológicas de sentido común [*commonsense psychological generalizations*] relacionan estados mentales en virtud de sus contenidos, y la representación canónica hace lo que puede para reconstruir tales relaciones de contenido como relaciones de forma. La "condición de Aristóteles" exige que nuestra teoría de las actitudes proposicionales racionalice ese proceso interpretando a los verbos de APs de manera de permitir la referencia a la forma de sus objetos. Hacer esto es legitimizar los supuestos de la psicología de sentido común [*commonsense psychology*], y, con ello, tam-

bién, los de la psicología real (es decir, la psicología cognitiva) (véase Fodor, 1975).

De hecho, podemos enunciar (y satisfacer) la condición de Aristóteles en una versión aun más fuerte. Digamos que algo es una *oración de creencia* si es de la forma *a cree (que) S*. Definamos la *fórmula correspondiente* [*the correspondent*] a una oración como la fórmula que consiste en la oración *S* por sí sola.[8] Antes señalé que la relación existente entre las condiciones de verdad de la creencia que una oración de creencia adscribe y las condiciones de verdad de la fórmula correspondiente a la oración de creencia es la siguiente: la creencia es verdadera si y sólo si la fórmula correspondiente lo es. Presumiblemente, esto es parte de lo que involucra considerar a la fórmula correspondiente a una oración de creencia como *expresando* la creencia adscripta.

Por consiguiente, no debería sorprender que nuestras intuiciones acerca de la forma de la creencia adscripta por una cierta oración de creencia estén determinadas por la forma lógica de la fórmula correspondiente. De este modo, la creencia de Juan de que María y Guillermo se van es, intuitivamente, una creencia conjuntiva (cf. la forma lógica de 'María y Guillermo se van'); la creencia de Juan de que Alfredo es un cisne blanco, es una creencia singular (cf. la forma lógica de 'Alfredo es un cisne blanco'), etcétera. Es esencial en estos ejemplos que entendamos 'creencia' *de manera opaca*; de lo contrario, la creencia de que *P* tendrá la forma lógica de cualquier oración equivalente a *P*. Pero así es como debe ser: es en virtud de su contenido *opaco* que la creencia de Juan de que *P* juega un rol sistemático en la vida mental de Juan: por ejemplo, en la determinación de sus acciones y en la causación de otros estados mentales. Por ello, es la interpretación opaca la que opera en modelos explicativos tales como el silogismo práctico y sus herederos espirituales.

Ahora puedo enunciar la condición de Aristóteles en su versión más fuerte (y final). Una teoría de las actitudes proposicionales debería legitimizar la adscripción de forma a los objetos de las actitudes proposicionales. En particular, debería explicar por qué la forma de una cre-

8. La tarea de definir 'fórmula correspondiente' se complica cuando los verbos de las APs toman como sus objetos oraciones *transformadas*, pero no necesitamos ocuparnos aquí de los detalles técnicos. Baste con afirmar que queremos que la fórmula correspondiente a 'Juan quiere irse' sea 'Juan se va'; la correspondiente a 'Juan objeta que María y Guillermo sean electos', sea 'María y Guillermo son electos', etcétera.

encia es idéntica a la forma lógica de la fórmula correspondiente a la oración que (opacamente) adscribe esa creencia.[9]

Una digresión. Uno puede sentirse tentado a argumentar que satisfacer la condición de Aristóteles es incompatible con satisfacer la condición de Frege; que la opacidad de las oraciones de creencia muestra la futilidad de asignar forma lógica a sus objetos. El argumento podría ser el siguiente. Las oraciones tienen forma lógica en virtud de su comportamiento respecto de las transformaciones lógicas; la forma lógica de una oración es el aspecto de su estructura que provee un dominio para esas transformaciones. Pero Frege nos ha mostrado que los objetos de los verbos de actitud proposicional son inferencialmente inertes. Por consiguiente, es una especie de broma hablar de la forma lógica de los objetos de las APs; ¿qué fuerza puede tener decir que una oración tiene la forma *P & Q* si uno también tiene que decir que la simplificación de la conjunción no es aplicable?

Es posible que un argumento tal motive la tesis de la fusión. Se trata, en el mejor de los casos, de un equívoco. En particular, borra la distinción entre lo que es implicado por lo que se cree, y lo que es implicado por creer lo que se cree. De manera menos críptica: si Juan cree *P & Q* entonces lo que Juan cree implica *P* y lo que Juan cree implica *Q*. Esto es indiscutible; *P & Q* es lo que Juan cree, y *P & Q* implica *P*, *Q*. Sería errado, pues, expresar la condición de Frege como "*P & Q* es semánticamente inerte en el contexto 'Juan cree...' "; dado que esto parece insinuar que *P & Q* implica *P* sólo de vez en cuando. (Otro argumento igualmente malo: *P & Q* a veces no implica *P*, a saber, cuando está bajo el alcance del operador 'no'.) Lo que cae bajo la condición de Frege no es, entonces, la oración que expresa lo que Juan cree (esto es, *P & Q*) sino la oración que expresa que Juan cree lo que cree (esto es, la oración 'Juan cree que *P & Q*'). Adviértase que la inercia de la última oración no es una excepción a la ley de simplificación de la conjunción, dado que la simplificación de la conjunción no ha sido definida para oraciones de la forma *a cree que P & Q* sino sólo para oraciones de la forma *P & Q*.

"Sin embargo", alguien podría afirmar, "si la forma lingüística #*P*

9. Estoy suponiendo que dos oraciones cuyas fórmulas correspondientes tienen una forma lógico-sintáctica *distinta*, no pueden asignar la *misma* creencia (opaca), y alguien podría cuestionar esto; considérese 'Juan cree que María engañó a Guillermo' y 'Juan cree que Guillermo fue engañado por María'. Este tipo de objeción es serio, y será respondido más adelante.

& Q# es lógicamente inerte en el contexto 'Juan cree...', ¿qué *sentido* tiene hablar de la forma lógica del complemento de las oraciones de creencia?". Esto no es, por supuesto, un argumento, pero es una pregunta relevante. Respuesta: (a) porque podemos querer satisfacer la condición de Aristóteles (por ejemplo, para estar en condiciones de enunciar el silogismo práctico); (b) porque podemos querer comparar creencias respecto de su forma (la creencia de Juan de que todos los *F*s son *G*s es una generalización de la creencia de María de que *a* es *F* y *G*; la creencia de Samuel de que *P* es incompatible con la creencia de Guillermo de que *no-P*; etcétera); (c) porque podemos querer hablar de las consecuencias de una creencia, aun cuando admitamos alegremente que las consecuencias de una creencia pueden no ser ellas mismas objeto de creencia (esto es, creídas). En realidad, necesitamos la noción de consecuencia de una creencia sólo si queremos decir que la creencia no es cerrada [*closed*] respecto de la relación de consecuencia.

Hasta aquí la digresión.

V. Una teoría de las actitudes proposicionales debería asociarse con explicaciones empíricas de los procesos mentales.

Queremos que una teoría de las APs nos diga qué *son* las actitudes proposicionales (en tanto casos [*tokens*]); o, al menos, en virtud de qué hechos son verdaderas las adscripciones de actitudes proposicionales. Me parece evidente que una teoría tal no podría ser aceptable a menos que diera cabida a la explicación de datos —gruesos y de sentido común o sutiles y experimentales— acerca de los procesos y estados mentales. Esto no implica exigir, por supuesto, que la teoría de las APs legitimice nuestra psicología empírica corriente; sólo exige que sea compatible con alguna psicología independientemente garantizada. Considero que esto es análogo a lo siguiente: la teoría de que el agua es H_2O no podría ser aceptable a menos que, junto con las adecuadas premisas empíricas, proporcionara una explicación de las macro y micro-propiedades del agua. Considero esto innegable.

Pienso, de hecho, que el requisito de que una teoría de las actitudes proposicionales deba ser empíricamente plausible, puede cumplir varias funciones; muchas más de lo que los filósofos normalmente creen. Volveré sobre esto más tarde, cuando tengamos algunas teorías a mano.

Éstas son pues las condiciones que quiero que cumpla una teoría de las actitudes proposicionales. Argumentaré que todas juntas sugieren fuertemente que las actitudes proposicionales son relaciones entre organismos y fórmulas en un lenguaje interno; entre organismos y oraciones

internas, por así decir. Es conveniente, sin embargo, dar los argumentos en dos pasos. Primero, mostrar que las condiciones I-V concuerdan con el punto de vista de que los objetos de las APs son oraciones; luego, mostrar que es plausible considerar que esas oraciones son internas.

Empiezo por anticipar la acusación de falsa propaganda. Los argumentos a los que pasaré revista, son explícitamente no-demostrativos. Todo lo que alego en favor de la teoría del lenguaje interno es que funciona (a) sorprendentemente bien, y (b) *mejor* que cualquiera de las alternativas disponibles. El punto decisivo viene al final: aun cuando no necesitáramos la tesis de la oración interna para los propósitos de I-V, la necesitaríamos para nuestra psicología. Sin duda, otro argumento no-demostrativo, pero un argumento que encuentro extremadamente persuasivo.

La teoría de Carnap

En *Meaning and Necessity* (1947), Rudolf Carnap sugirió que las APs pueden ser interpretadas como relaciones entre personas y las oraciones que esas personas están dispuestas a emitir; por ejemplo, entre personas y oraciones del castellano. Carnap tenía en mente hacer frente al problema de la opacidad, pero es sorprendente e instructivo encontrar que su propuesta satisface bastante bien *todas* las condiciones enumeradas. Considérese lo siguiente:

I. Si las actitudes proposicionales son relaciones con oraciones, entonces son relaciones *tout court*. Más aún, se asume que la relación adscripta por una oración de la forma *a cree...* vale entre el individuo denotado por '*a*' y la fórmula correspondiente a la cláusula complementaria. Es, pues, claro por qué la creencia adscripta a *a* es verdadera si y sólo si la fórmula correspondiente lo es; si la teoría de Carnap es correcta, la fórmula correspondiente es el *objeto* de la creencia (esto es, la fórmula correspondiente es lo que es creído-como-verdadero [*what's believed-true*]).

II. Es probable que la condición de Vendler se pueda satisfacer, si bien los detalles dependerán de cómo se interprete a los objetos de los verbos de decir. La estrategia natural de un neo-carnapiano sería considerar que 'Juan dijo que *P*' es verdadera en virtud de alguna relación entre Juan y un ejemplar del tipo *P*. Dado que, según esta explicación,

decir *P* y creer *P* involucran relaciones con ejemplares de la misma oración, es poco sorprendente que las fórmulas que expresan el objeto de la relación *dice-que* resulten ser similares, lógica y sintácticamente, a las fórmulas que expresan el objeto de la relación *cree-que*.

III. La condición de Frege es satisfecha; la opacidad de la creencia es interpretada como un caso especial de la opacidad de la oración citada. Con otras palabras, "Juan dijo 'Guillermo engañó a María' " expresa una relación entre Juan y una oración (citada [*quoted*]), de manera que el hecho de que Juan pueda tener *esa* relación con *esa* oración, sin tenerla con alguna otra arbitrariamente similar pero distinta, tal como 'Alguien engañó a María' o 'Guillermo engañó a alguien', no nos resulta sorprendente. Pero de manera similar, mutatis mutandis, 'Juan cree que Guillermo engañó a María' *también* expresa una relación entre Juan y una oración citada.

IV. La condición de Aristóteles es satisfecha de manera total. La forma lógica del objeto de una oración de creencia es heredada de la forma lógica de la fórmula correspondiente a la oración de creencia. Esto es obvio, dado que, en la concepción de Carnap, la fórmula correspondiente a la oración de creencia *es* el objeto de la creencia que ella adscribe.

V. La plausibilidad empírica que se asigne a la teoría de Carnap depende de cuáles sean los hechos empíricos que se reconocen acerca de las actitudes proposicionales, y de cuán ingenioso se sea para aprovechar la teoría de manera de proporcionar explicaciones de esos hechos. He aquí un ejemplo de cómo podría ser una explicación tal.

Es plausible alegar que existe un paralelismo bastante general entre la complejidad de las creencias y la complejidad de las oraciones que las expresan. Considero pues que, por ejemplo, 'La segunda Guerra Púnica fue librada bajo condiciones tales que ninguno de los combatientes podría haber deseado o previsto' es una oración más compleja que, por ejemplo, 'Está lloviendo'; y, correlativamente, considero que el pensamiento de que la Segunda Guerra Púnica fue librada bajo condiciones tales que ninguno de los combatientes podría haber deseado o previsto, es un pensamiento más complejo que el pensamiento de que está llo-

viendo. La teoría de Carnap explica este paralelismo,[10] dado que, de acuerdo con ella, lo que hace verdadera a una adscripción de creencia es la relación entre un organismo y la fórmula correspondiente a la oración adscriptora-de-creencia [*belief-ascribing sentence*]. De tal modo, tener la creencia de que la Segunda Guerra Púnica... etcétera es estar relacionado con una oración más compleja que aquella con la que se está relacionado cuando uno cree que está lloviendo.

Algunas personas necesitan contar narices antes de admitir que tienen una. En tal caso, véase el debate sobre "codificabilidad" ["*codability*"] en Brown y Lenneberg (1954) y en Brown (1976). Los experimentos muestran que la complejidad relativa de las descripciones que los sujetos proporcionan frente a trazos de colores predice la dificultad relativa que tienen en identificar los trazos en la etapa de recuerdo y reconocimiento. Brown y Lenneberg explican sus hallazgos siguiendo (inadvertidamente) lineamientos carnapianos estrictos: descripciones complejas corresponden a recuerdos complejos porque es la descripción lo que el sujeto recuerda (opacamente) cuando recuerda (transparentemente) el color del trazo.

Podemos comenzar a ver ahora *una* de las maneras en las cuales se supone que funciona la condición V. Una teoría de las actitudes proposicionales especifica una interpretación de los objetos de las actitudes. Dice algo en favor de esa teoría el que pueda demostrarse que se combina con un planteo plausible, por razones independientes, del "cálculo de costo" ["*cost-accounting*"] de los procesos mentales. Una función de cálculo de costo es un ordenamiento (parcial) de los estados mentales de acuerdo con su complejidad relativa. Dicho ordenamiento responde, a su vez, a una variedad de tipos de datos empíricos, tanto intuitivos como experimentales. *Grosso modo*, se logra una "conexión" entre un cálculo de costo empíricamente garantizado y una teoría de los objetos de las APs cuando se puede predecir la complejidad relativa de un estado (o proceso) mental a partir de la complejidad relativa de aquello que la teoría le asigna como su objeto (o dominio). (Por lo tanto, si Carnap está en lo cierto, la complejidad relativa de las creencias debería poder predecirse a partir de la complejidad lingüística relativa de las fórmulas correspondientes a las oraciones adscriptoras-de-creencia, en igualdad de circunstancias [*all other things being equal*].)

Hay más que decir acerca de esto de lo que el espacio me permite.

10. Al hablar de la teoría de Carnap, no deseo implicar que Carnap habría aceptado las aplicaciones que le estoy dando; todo lo contrario, me imagino.

Nuevamente, *grosso modo*: exigir que la complejidad de los objetos putativos de las APs prediga el cálculo de costo de las actitudes es imponer restricciones empíricas a la *notación* de las oraciones (canónicas) adscriptoras-de-creencias. Por lo tanto, si consideramos que el objeto de una AP es la fórmula correspondiente a la oración adscriptora-de-creencia obtendremos claramente predicciones acerca de la complejidad relativa de las creencias, distintas de las que obtendríamos si consideráramos que el objeto es, por ejemplo, la fórmula correspondiente transformada en la forma disyuntiva. El hecho de que haya consecuencias empíricas de la notación que usamos para especificar los objetos de las APs forma parte, por supuesto, del hecho de que interpretamos las adscripciones de actitud *de manera opaca*; es precisamente bajo una interpretación opaca que distinguimos (por ejemplo) el estado mental de creer que *P & Q* del estado mental de creer que ni no-*P* ni no-*Q*.

En síntesis, a la teoría de Carnap le va bastante bien con las condiciones I-V; hay más para decir en su favor de lo que podría inferirse del callado entusiasmo con que los filósofos, en general, la han recibido. Sin embargo, pienso que el consenso filosófico está justificado; la teoría de Carnap no tiene éxito. Éstas son algunas de las razones.

1. Carnap tiene una teoría acerca de los objetos de las actitudes proposicionales (es decir, son oraciones) y una teoría acerca del carácter de la relación con esos objetos, en virtud de la cual uno tiene una creencia, un deseo, etcétera. Ahora bien, esta última teoría es manifiestamente conductista. Según la concepción de Carnap, creer que tal o cual cosa es estar dispuesto a emitir (en condiciones presumiblemente especificables) ejemplares de la fórmula correspondiente a la oración adscriptora-de-creencia. Pero es evidente que las creencias no son disposiciones a actuar; a fortiori, no son disposiciones a emitir nada. Por consiguiente, al menos parte de la explicación carnapiana de las actitudes es errónea.

He planteado esta objeción en primer término porque es la más fácil de encarar. Hasta donde yo puedo ver, nada impide a Carnap mantener su explicación de los *objetos* de creencia y renunciar al análisis conductista de la relación de creencia. Esto lo movería a buscar nuevas respuestas a preguntas tales como: ¿cuál es la relación con la oración 'Está lloviendo' tal que uno cree que está lloviendo si y sólo si uno está en esa relación? En particular, lo movería a buscar una respuesta que no fuese la conductista: "Es la relación de estar dispuesto a emitir ejemplares de esa oración cuando...".

La solución natural para Carnap sería volverse funcionalista; soste-

ner que creer que está lloviendo es hacer que un ejemplar de 'Está lloviendo' desempeñe un cierto rol en la causación de la conducta y en la de (otros) estados mentales, rol que debería ser especificado eventualmente en el curso de una elaboración detallada de la psicología empírica..., etcétera, etcétera. Esto tal vez no sea una tesis, pero está de moda, no conozco nada mejor, y tiene la virtud de explicar por qué las actitudes proposicionales son opacas. *Grosso modo*, sobre la base de una relación lógica cualquiera entre S_1 y S_2 (excepto, por supuesto, la de identidad), uno no esperaría poder inferir 'ejemplares de la oración S_2 tienen el rol causal *R*' a partir de 'ejemplares de la oración S_1 tienen el rol causal *R*'. Generalizando aún más, hasta donde yo puedo ver, una explicación funcionalista del modo en que las oraciones citadas figuran en la atribución de APs servirá respecto de las condiciones I-V tan bien como una explicación en términos de disposiciones a emitir. De ahora en adelante daré por sentada esta enmienda.

2. La manera natural de leer la teoría de Carnap es considerar a la identidad de tipo de las fórmulas correspondientes a las oraciones adscriptoras-de-creencia como (condición) necesaria y suficiente para la identidad de tipo de las creencias adscriptas, y es posible argüir, al menos, que esto recorta demasiado finamente a las APs. De este modo, uno podría sostener plausiblemente, por ejemplo, que 'Juan cree que María engañó a Guillermo' y 'Juan cree que Guillermo fue engañado por María' adscriben la misma creencia.[11] En efecto, éste es el lado oscuro de la estrategia que consiste en derivar la opacidad de la creencia de la opacidad de la cita. La estrategia fracasa toda vez que las condiciones de identidad de las creencias son *distintas* de las condiciones de identidad de las oraciones.

Una manera de solucionar el problema sería concebir a los objetos de las creencias como *conjuntos intertraducibles* [*translation sets*] de oraciones; una idea similar parece motivar la doctrina carnapiana del isomorfismo intencional. De todos modos, los problemas en esta área son conocidos. Bien podría ser, por ejemplo, que la manera correcta de caracterizar una relación de traducción entre oraciones hiciera referencia a las intenciones comunicativas de los hablantes/oyentes del lenguaje al que las oraciones pertenecen, cualquiera que sea éste. (S_1 traduce S_2 si y sólo si las dos oraciones son usadas corrientemente con las mismas intenciones comunicativas.) Pero, por supuesto, no podemos al mismo

11. Véase la nota 9.

tiempo identificar traducciones por referencia a intenciones e individuar actitudes proposicionales (que incluyen intenciones) por referencia a traducciones. Este problema persiste independientemente de preocupaciones gnoseológicas acerca de la facticidad de las adscripciones de actitudes proposicionales, la determinación o indeterminación de las traducciones, etcétera; lo cual sugiere que podría ser serio.

3. Uno puede creer que está lloviendo aun cuando no hable castellano. Ésta es una variante del problema de la fineza de análisis antes mencionado; sugiere, una vez más, que los objetos de creencia adecuados son conjuntos de traducciones, y reaviva el espectro que amenaza a ese enfoque.

4. Ciertamente, uno puede creer que está lloviendo aun cuando uno no hable ningún lenguaje. Decir esto es decir que al menos *alguna* psicología humana cognitiva es generalizable a organismos infrahumanos; si no fuera así, encontraríamos la conducta de los animales *completamente* desatinada, lo cual, de hecho, no es el caso.

Por supuesto, hablar de relaciones es fácil; debe haber *alguna* relación entre un perro y 'Está lloviendo' si y sólo si el perro cree que está lloviendo; aunque, quizás, una relación no muy interesante. ¿Por qué elegirla entonces como la relación en virtud de la cual la adscripción de creencia vale para el perro? El problema es la condición V. Sería simplemente un milagro si hubiera una relación entre perros y ejemplares de 'Está lloviendo' tal que alguno de los hechos empíricos acerca de la atribución de actitudes proposicionales a los perros fuera explicable en términos de dicha relación. (No podemos, por ejemplo, elegir ninguna relación funcional/causal porque, sin duda, la conducta de los perros no es causada por ejemplares de oraciones del castellano.) Para expresarlo de manera general, si bien un poco cruda: satisfacer la condición V implica suponer que lo que la teoría considera como el objeto de una AP debe jugar un rol adecuado en los procesos mentales del organismo al cual la actitud es adscripta. Pero las oraciones del castellano no desempeñan ningún rol en la vida mental de los perros. (Exceptuando, tal vez, oraciones como '¡Siéntate, Rover!', las cuales nunca desempeñan el tipo de rol aquí considerado.)

5. Argumenté que las creencias heredan sus condiciones de verdad de las fórmulas correspondientes a las oraciones adscriptoras-de-creencias; pero esto no funciona si, por ejemplo, existen creencias inexpresables. Este problema es especialmente grave para las explicaciones con-

ductistas (o funcionalistas) de la relación de creencia; creer que *P* no puede ser cuestión de estar dispuesto a emitir (o de que nuestra conducta esté causada por) ejemplares de la oración *P* si, de hecho, no hay tal oración. Sin embargo, es el apelar a las oraciones citadas lo que resuelve este punto en tales teorías: lo que les permite satisfacer I-V.

6. Señalé que hay una correspondencia aproximada entre la complejidad de los pensamientos y la complejidad de las oraciones que los expresan, y que la teoría (neo-)carnapiana tiene en cuenta esto; en general que la concepción de que los objetos de las APs son oraciones del lenguaje natural, podría entrelazarse razonablemente bien con un cálculo de costo de los estados y procesos mentales, empíricamente defendible. Lamentablemente, este argumento tiene un doble filo si suponemos —como parece ser plausible— que la correspondencia es sólo parcial. Toda vez que falle, habrá evidencia prima facie *en contra de* la teoría de que las oraciones son los objetos de las actitudes proposicionales.

De hecho, en este punto, podemos hacer algo mejor que apelar a intuiciones. A modo de ejemplo: más arriba, destaqué que la "codificabilidad" (esto es, la simplicidad media de las descripciones en castellano) de los colores predice su evocación en una población de hablantes del castellano, y que esto concuerda con el punto de vista según el cual lo que uno recuerda, cuando recuerda un color, es (al menos a veces) su descripción: es decir, concuerda con el punto de vista de que las descripciones son los objetos de (al menos algunas) actitudes proposicionales. Es, pues, sorprendente encontrar que la codificabilidad *en castellano* también predice la evocación en una población de hablantes del dani. No podemos explicar esto presuponiendo una correlación entre la codificabilidad-en-castellano y la codificabilidad-en-dani (esto es, suponiendo que los colores que los hablantes del castellano encuentran fáciles de describir son los que los hablantes del dani también encuentran fáciles de describir) dado que, según se ha observado, el dani no tiene *ningún* tipo de vocabulario para la variación cromática; tales variaciones son *totalmente* incodificables en dani. Esto se parece a la temida paradoja señalada antes: ¿cómo podrían las oraciones del *castellano* ser objeto de las actitudes proposicionales de hablantes (monolingües) del dani? Y, si no lo son, ¿cómo podría una propiedad definida para las oraciones del castellano encajar en una teoría del cálculo de costo de los procesos mentales de los danis? Parecería que o bien (a) algunas actitudes proposicionales *no* son relaciones con oraciones, o bien (b) si lo son —si las

oraciones del castellano son de alguna manera los objetos de las APs de los danis— entonces las oraciones que constituyen los objetos de las APs no necesitan desempeñar ningún rol funcional/causal en la posesión de las actitudes. (Para una discusión de los resultados interculturales relativos a la codificabilidad, véase Brown, 1976. Para los detalles de los estudios originales, véase Heider, 1972 y Berlin y Kay, 1969.)

7. Si las oraciones (ejemplar) de un lenguaje natural son los objetos de las actitudes proposicionales, ¿cómo se aprende el lenguaje (materno)? En toda teoría del aprendizaje del lenguaje que podamos concebir, dicho proceso tiene que involucrar la recolección de datos, la formulación de hipótesis, la contrastación de las hipótesis con los datos, y una decisión respecto de qué hipótesis es mejor confirmada por los datos. Con otras palabras, debe involucrar estados y procesos mentales tales como creencias, expectativas e integración perceptual. Es importante darse cuenta de que nadie ha propuesto jamás explicación *alguna* del aprendizaje lingüístico que no involucrara actitudes proposicionales y estados mentales, con la sola excepción de los conductistas. Y las explicaciones conductistas del aprendizaje lingüístico son, sin duda, insostenibles. Por consiguiente, si se quiere evitar la circularidad, debe admitirse que *algunas* actitudes proposicionales no son relaciones funcionales/causales con oraciones del lenguaje natural. No veo manera de evitar esto que no sea una opción peor que el rechazo de la teoría de Carnap.

La situación es, entonces, desalentadora. De un lado, contamos con argumentos plausibles a favor de la teoría de Carnap (a saber, I-V), y del otro, disponemos de argumentos igualmente plausibles en su contra (a saber, 1-7). No importa; al reconsiderar la situación, parecería que no necesitamos aceptar toda la teoría de Carnap para satisfacer I-V, ni necesitamos rechazarla toda para evitar 1-7. *Grosso modo*, todo lo que I-V requieren es la parte de la teoría que dice que los objetos de las APs son *oraciones* (y tienen por lo tanto, formas lógicas, condiciones de verdad, etcétera). Mientras que lo que produce los problemas señalados en 1-7 es la parte de la teoría que dice que son oraciones del *lenguaje natural* (surgiendo entonces los problemas relacionados con los organismos no-verbales, el aprendizaje del lenguaje materno, etcétera) la solución recomendable es, entonces, considerar a los objetos de las APs como oraciones de un lenguaje *no-natural*: como fórmulas de un Sistema Representacional Interno [SRI].

El primer punto es demostrar que esta propuesta logra lo que se

presume que logra: satisface I-V sin plantear problemas respecto de 1-7. De hecho, propongo hacer menos que eso, dado que, según creo, los detalles serían extremadamente complejos. Baste entonces con indicar aquí la estrategia general.

Las condiciones I y III son relativamente fáciles de satisfacer. La condición I exige que las actitudes proposicionales sean relaciones, y sin duda lo son si es que son relaciones con representaciones internas. La condición III exige una interpretación de la opacidad. Carnap cumplió con esta exigencia al reducir la opacidad de la creencia a la opacidad de la cita, y lo mismo haremos nosotros; la única diferencia es que, mientras que para Carnap 'Juan cree que está lloviendo' relaciona a Juan con una oración del castellano, para nosotros lo relaciona con una fórmula interna.

Las condiciones II y IV enfatizan el paralelismo lógico-sintáctico entre los complementos y las fórmulas correspondientes a las oraciones que adscriben-creencias; dichas relaciones son compendiadas en la identidad entre las condiciones de verdad de 'Está lloviendo' y las de lo que se cree cuando se cree que está lloviendo. La teoría (neo-)carnapiana explicaba esas simetrías considerando a las fórmulas correspondientes a las adscripciones-de-creencia como los objetos mismos de creencia. La presente alternativa es similar en espíritu pero menos directa: suponemos que la fórmula correspondiente a una oración adscriptora-de-creencia hereda sus propiedades lógico-semánticas de la fórmula interna que funciona como el objeto de la creencia adscripta.

Hay tres piezas en juego: (a) los *adscriptores-de-creencia* (como 'Juan cree que está lloviendo'); (b) los *complementos* [*complements*] de adscriptores-de-creencia (como la frase 'está lloviendo' en 'Juan cree que está lloviendo'), y (c) las *fórmulas correspondientes* a los adscriptores-de-creencia (como la oración no subordinada 'Está lloviendo'). La idea es hacer que las tres converjan (aunque, por supuesto, por caminos distintos) en la misma fórmula interna (llámesela '*F* (está lloviendo)'),[12] proporcionando de esta manera el fundamento (teórico) para explicar las analogías expresadas por II y IV. Para que esto funcione adecuadamente, habría que dar instrucciones detalladas para conectar la teoría de las APs con la teoría de la interpretación de oraciones, y he ubicado mal a la mía, pero la idea general es clara. Los adscriptores-de-creencia son verdaderos en virtud de relaciones funcionales/causales (llámeselas

12. Donde *F* podría concebirse como una función, que va de oraciones (por ejemplo, del castellano) a fórmulas internas.

'generadoras-de-creencia' [*belief-making*]) entre organismos y casos de fórmulas internas. Así, en particular, 'Juan cree que está lloviendo' es verdadera en virtud de una relación generadora-de-creencia entre Juan y un caso de *F* (está lloviendo). Es, por supuesto, el complemento de un adscriptor-de-creencia el que determina *qué* fórmula interna está involucrada en las condiciones de verdad; en efecto, 'está lloviendo' en 'Juan cree que está lloviendo' funciona como un índice que identifica *F* (está lloviendo) (en vez de, por ejemplo, *F* (los elefantes tienen alas)), como la fórmula interna con la cual Juan está relacionado si y sólo si 'Juan cree que está lloviendo' es verdadera.

Por lo tanto, visto desde cierta perspectiva, el complemento conecta a un adscriptor-de-creencia con una fórmula interna. Pero, simultáneamente, también lo conecta con la fórmula correspondiente: es porque la expresión 'está lloviendo' constituye el complemento del adscriptor-de-creencia que la fórmula correspondiente a 'Juan cree que está lloviendo' es 'Está lloviendo'. Y ahora podemos cerrar el círculo, dado que, por supuesto, *F* (está lloviendo) está *también* semánticamente conectada con la fórmula correspondiente a 'Juan cree que está lloviendo', esto es, por el principio según el cual 'Está lloviendo' es la oración que los hablantes del castellano usan cuando se encuentran en una relación generadora-de-creencia con un caso de *F* (está lloviendo) y quieren usar una oración del castellano para decir lo que creen.

Hay varias maneras de concebir la relación entre las fórmulas internas y las fórmulas correspondientes a los adscriptores-de-creencia. Una es pensar que las convenciones de un lenguaje natural funcionan estableciendo un apareamiento entre sus formas verbales y las fórmulas internas que mediatizan las actitudes proposicionales de sus usuarios; en particular, los objetos internos de creencia aparecen con la forma lingüística que los hablantes/oyentes usan para expresar sus creencias. Ésta es una manera natural de enfocar la situación si se concibe al lenguaje natural como un sistema de vehículos convencionales para la expresión del pensamiento (concepción a la que no le encuentro ningún defecto grave). Así, en nuestro ejemplo, las convenciones del castellano asocian: 'Está lloviendo' con *F* (está lloviendo) (esto es, con el objeto de la creencia de que está lloviendo); 'Los elefantes tienen alas' con *F* (los elefantes tienen alas) (esto es, con el objeto de la creencia de que los elefantes tienen alas); y, en general, el objeto de cada creencia con la fórmula correspondiente a una oración adscriptora-de-creencia.[13]

13. Suponiendo —como podríamos hacer, aunque no es ahora necesario— que todas

Otra opción es suponer que *F* (está lloviendo) se distingue por el hecho de que sus casos desempeñan un rol causal/funcional (no sólo como objeto de la creencia de que está lloviendo, sino también) en la producción de emisiones lingüísticamente regulares de 'Está lloviendo'. En realidad, esta opción podría usarse en tándem con la anterior, dado que sería razonable interpretar a las emisiones "lingüísticamente regulares" como aquéllas que son producidas a la luz del conocimiento que el hablante tiene de las convenciones lingüísticas. De cualquier manera, la idea básica sería considerar a *F* (está lloviendo) como el objeto de las intenciones comunicativas que expresan normalmente las emisiones de 'Está lloviendo'; por lo tanto, como una de las causas *mentales* de tales emisiones. Considero que, dada esta relación, debe ser posible desarrollar tácticas detalladas para satisfacer las condiciones II y IV; pero ésta es la parte que me propongo dejar librada a la inventiva del lector. Lo que quiero enfatizar aquí es la manera en que la estructura lingüística del complemento de un adscriptor-de-creencia lo conecta (en una dirección) con oraciones declarativas no subordinadas y (en otra) con fórmulas internas. A diferencia de la tesis de la fusión, no es accidental que la expresión 'está lloviendo' ocurra en 'Juan cree que está lloviendo'. La facilidad de los lenguajes naturales para decir lo que uno cree *y* que uno lo cree se manifiesta en el aprovechamiento de esta elegante simetría.

¿Qué pasa con la condición V? La voy a considerar en conjunción con 2-7, dado que lo que merece destacarse acerca de éstas últimas es que todas ellas registran objeciones *empíricas* contra la explicación carnapiana. Por ejemplo, 3, 4 y 6 no tendrían fuerza alguna si todos (es decir, todo sujeto de adscripciones verdaderas de actitudes proposicionales) hablaran castellano; 2 y 5 dependen de la probabilidad empírica de que las oraciones del castellano no se correspondan unívocamente con los objetos de las actitudes proposicionales; 7 sería satisfecha si el castellano fuera innato. En realidad, supongo que un neo-carnapiano de la línea ultra dura podría considerar que es posible salvar la situación afirmando que —pese a las apariencias— el castellano es innato, universal, lo suficientemente rico, etcétera. Mi punto es que éste es el *tipo* de movida adecuado; lo que se puede decir en su contra es que es palpablemente falso.

Por lo demás, parte del encanto de la idea de un lenguaje interno

---

las creencias puedan ser expresadas en castellano. Es, por supuesto, una consecuencia de la presente propuesta el que todas las creencias que *podamos* tener sean expresables en el código interno.

es que, dado que no se sabe prácticamente nada acerca de los detalles de los procesos cognitivos, podemos hacer las suposiciones que correspondan acerca del sistema representacional interno, arriesgando en el peor de los casos nada más que una crasa implausibilidad.

Por lo tanto, supongamos —lo cual, en todo caso, no *sabemos* que sea falso— que el lenguaje interno es innato, que sus fórmulas se corresponden unívocamente con los contenidos de las actitudes proposicionales (por ejemplo, que 'Juan engañó a María' y 'María fue engañada por Juan' se corresponden con la misma "oración interna"), y que es *tan* universal como la psicología humana; esto es, que en la medida en que un organismo comparta nuestros procesos mentales, también compartirá nuestro sistema de representaciones internas. Dadas estas suposiciones, todo encaja bien. Ya no resulta paradójico, por ejemplo, que la codificabilidad *en castellano* prediga la relativa complejidad de los procesos mentales de los danis; puesto que, de acuerdo con nuestras suposiciones, no es *realmente* la complejidad de las oraciones del castellano lo que *nuestro* cálculo de costo predice; no esperamos que *esa* correspondencia sea más que parcial (véase la objeción 6). Lo que nuestro cálculo de costo realmente predice es la complejidad relativa de las representaciones internas que expresamos mediante el uso de oraciones del castellano. Y, nuevamente en virtud de nuestros supuestos, el sistema subyacente de representaciones internas es común a los danis y a nosotros. Si este supuesto no gusta, traten de encontrar alguna otra hipótesis que explique los hechos acerca de los danis.

Adviértase que decir que hacemos suposiciones empíricas no equivale a decir que las hacemos gratuitamente. Ellas conllevan un cuerpo de compromisos empíricos que de ser insostenibles, frustrarían la idea de representación interna. Imagínese, por ejemplo, que el cálculo de costo para los hablantes del castellano demuestra no estar relacionado con el cálculo de costo para (por ejemplo) los hablantes del latvio. (Imagínese, en efecto, que la hipótesis de Whorf-Sapir resulta ser más o menos verdadera.) Resulta entonces difícil ver cómo es posible que el sistema de representaciones internas sea universal. Pero si no es universal, es de presumir que no sea innato. Y si no es innato, no puede mediar en el aprendizaje de los lenguajes maternos. Y si no puede mediar en el aprendizaje de los lenguajes maternos, carecemos de medios para responder a la objeción 7. Hay muchas maneras de poder averiguar que la teoría es errónea si, de hecho, lo es.

A lo que hemos llegado es a esto: las características generales de las actitudes proposicionales parecen requerir como objeto entidades de

tipo oracional. Y condiciones empíricas generales parecen impedirnos identificar esas entidades con las oraciones de los lenguajes *naturales*; de ahí que postulemos representaciones internas y lenguajes privados. ¿Qué mal hay en haber llegado a esto? Quiero argumentar ahora que la actual conclusión es requerida en forma independiente, por cuanto es presupuesta por la mejor —en realidad, la única— psicología que hemos conseguido. No es sólo, como algún filósofo ha señalado con cierta irresponsabilidad, que "a algunos psicólogos les guste hablar así", sino que las mejores explicaciones de los procesos mentales con que contamos son totalmente ininteligibles a menos que algo parecido a la teoría de la representación interna sea verdadero.

La manera sistemática de defender este punto es por medio de una discusión detallada de tales teorías, pero me he dedicado a ello en otra parte, y es suficiente. Baste con considerar aquí un único ejemplo que es, sin embargo, prototípico. Reitero, una vez más, que los detalles no importan; que uno podría defender los mismos puntos considerando fenómenos extraídos de cualquier área de la psicología cognitiva lo suficientemente desarrollada como para garantizar que se hable de una teoría *in situ*.

Por lo tanto, tómese un fragmento de la (psico)lingüística contemporánea; considérese la explicación que se da de la ambigüedad de una oración como '*They are flying planes*' (a partir de aquí, a menudo, *S*). La explicación convencional es la siguiente: la oración es ambigua porque hay dos maneras de agrupar en frases a la secuencia de palabras, dos maneras de utilizar los "paréntesis". Una manera, que surge de interpretar a la oración como respondiendo a la pregunta "¿Qué son esas cosas?", es la siguiente: (*They*) (*are*) (*flying planes*). Es decir, la oración es copulativa, el verbo principal es '*are*' y '*flying*' es un adjetivo que modifica a '*planes*'. Pero, según la otra manera de agrupar las palabras, que surge de interpretar a la oración como respondiendo a la pregunta "¿Qué está haciendo esa gente?", se obtiene: (*They*) (*are flying*) (*planes*); es decir, la oración es transitiva, el verbo principal es '*flying*' y '*are*' es el auxiliar. Supongo, sin argumentar, que algo parecido explica la ambigüedad de *S*, o al menos ayuda a explicarla. La evidencia en favor de este tipo de enfoque es abrumadora y no hay, literalmente, ninguna alternativa teórica en este campo.

Pero, ¿qué puede significar decir de *S* que "tiene" dos agrupaciones distintas? Vuelvo a andar un camino conocido: *S* tiene dos agrupaciones por cuanto existe una función (llámesela *G-propia*) [*G-proper*] que va de la palabra 'oración' a las secuencias de palabras entre paréntesis que

constituyen oraciones del inglés. Y tanto '(*They*) (*are*) (*flying planes*)' como '(*They*) (*are flying*) (*planes*)' pertenecen al dominio de esa función. (Más aun, ninguna otra agrupación de esa secuencia de palabras pertenece al dominio de *G-propia*..., etcétera.)

Ahora bien, el problema que genera esta explicación, tal como se ha formulado, es que o bien es entimemática o bien es tonta. Uno quiere saber, cómo *podría* la mera existencia, si se quiere, platónica, de *G-propia* explicar hechos acerca de la ambigüedad de las oraciones del inglés. O, para decirlo de otra manera, es seguro que existe, de manera platónica, una función según la cual *S* recibe dos agrupaciones. Pero también existe, platónicamente, una función *G*', según la cual recibe dieciséis, y una función *G*'' según la cual recibe siete, y una función *G*''' según la cual no recibe ninguna. Dado que *G*', *G*'' y *G*''', en tanto funciones, son tan adecuadas como *G-propia*, ¿cómo podría su mera *existencia* explicar las propiedades lingüísticas de *S*? Tal vez alguien se sienta inclinado a decir: "Ah, pero *G-propia* es la (o tal vez *la*) gramática del inglés, y eso es lo que la distingue de *G*', *G*'' y de todas las demás". Pero esta explicación es inconducente porque sugiere la pregunta: ¿por qué la gramática del inglés desempeña un rol especial en la explicación de las oraciones del inglés? O, para formular la misma pregunta aunque de manera levemente distinta: llámese a *G*' la schmamática [*schmammar*] del inglés. Queremos saber ahora cómo es que la agrupación asignada por la gramática inglesa, es la que predice la ambigüedad de '*They are flying planes*', y no la asignada por la schmamática inglesa.

En mi opinión, hay una sola manera de responder a estas preguntas; esto es, sostener que *G-propia* (no sólo existe sino que) es el sistema mismo de fórmulas (internas (¿qué otra posibilidad hay?)) que los hablantes/oyentes del inglés usan para representar las oraciones de su lenguaje. Pero, si aceptamos esto, nos comprometemos, querámoslo o no, a hablar de al menos *algunos* procesos mentales (procesos de comprensión y producción de oraciones) que incluyen al menos algunas relaciones con al menos algunas representaciones internas. Y, si de alguna manera hemos de tener representaciones internas, ¿por qué no considerarlas los objetos de las actitudes proposicionales, aplacando de esta manera I-V? Digo "si aceptamos esto", pero en realidad no tenemos alternativa. Puesto que hay buena evidencia a favor de esta explicación, la cual —por lo demás—, no puede probarse que sea incoherente, y, para repetirlo otra vez, es la única explicación disponible. Una ciencia que funciona adquiere ipso facto buena reputación filosófica.

De tal modo, a través de una serie de argumentos no-demostrativos:

existen representaciones internas, y las actitudes proposicionales son relaciones que tenemos con ellas. Me falta discutir dos objeciones estrechamente relacionadas.

Objeción 1: ¿Por qué no considerar que los objetos de las actitudes proposicionales son *proposiciones*?

Esta sugerencia tiene sin duda un dejo de plausibilidad etimológica. De hecho, hasta donde yo sé, podría ser correcta. El error está en suponer que de alguna manera entra en conflicto con la presente propuesta.

Tomo en serio la idea de que el sistema de representaciones internas constituye un lenguaje (computacional). En tanto lenguaje, tiene, presumiblemente, una sintaxis y una semántica; especificar el lenguaje implica decir en virtud de qué propiedades sus fórmulas están bien formadas y qué relación(es) existe(n) entre las fórmulas y las cosas en el mundo (no lingüístico). No tengo idea de cómo podría ser una semántica adecuada para un sistema de representaciones internas; baste con decir que si las proposiciones desempeñan algún rol, lo desempeñan aquí. En particular, nada nos impide especificar una semántica para el SRI diciendo (*inter alia*) que algunas de sus fórmulas expresan proposiciones. Si decimos esto, podemos darle sentido a la idea de que las actitudes proposicionales son relaciones con proposiciones; esto es, son relaciones *mediadas*, y las representaciones internas hacen de mediadoras.

Ésta es, en términos generales, la manera como funcionan las teorías representacionales de la mente. Así, en las versiones clásicas, pensar en Juan (interpretado de manera opaca) es una relación con una "idea", es decir, con una representación interna de Juan. Pero esto es perfectamente compatible con interpretarlo también (de manera transparente) como una relación con *Juan*. En particular, cuando García está pensando en Juan, (normalmente) está en relación con Juan y lo está *en virtud de* estar en relación con la idea de Juan. De la misma manera, mutatis mutandis, si pensar que va a llover es estar en relación con una proposición, entonces, según la explicación presente, se está en esa relación en virtud de una relación (funcional/causal) con una fórmula interna que expresa la proposición. Sin duda, "expresar" es un tanto oscura; pero ése es un problema que concierne a las proposiciones y no a las representaciones internas.

"Ah, pero si se va a admitir proposiciones como los objetos *mediados* de las actitudes proposicionales, ¿por qué preocuparse por las representaciones internas como sus objetos inmediatos? ¿Por qué no decir: 'Las actitudes proposicionales son relaciones con proposiciones. ¡Punto!' ". Hay una razón menor y una importante. La razón menor es

que las proposiciones no tienen las propiedades adecuadas para nuestros fines. En particular, uno anticipa problemas de cálculo de costo. Como se recordará, la condición V nos permite elegir entre distintas teorías de APs en virtud de la forma de las entidades que ellas asignan como objetos de las actitudes. Ahora bien, el problema con las proposiciones es que son tipos de cosas que, presumiblemente, no *tienen* forma. Las proposiciones son contenidos puros; neutralizan las diferencias léxico-sintácticas entre distintas maneras de decir la misma cosa. *Para eso* son. Digo que este problema es menor, pero se torna enorme si lo que se desea es una teoría del objeto de las APs que tenga reputación empírica. Después de todo, no es sólo el cálculo de costo lo que se supone determinado por los aspectos formales de los objetos de las APs; son *todos* los procesos y propiedades mentales explicados por la psicología cognitiva. Eso es lo que *significa* hablar de una psicología *computacional*. Los principios computacionales son principios que se aplican en virtud de la forma de las entidades de su dominio.

Pero la razón principal que tengo para no decir "Las actitudes proposicionales son relaciones con proposiciones. ¡Punto!" es que no lo entiendo. No entiendo cómo un organismo puede estar en una relación (cognoscitiva interesante) con una proposición, si no es manteniendo una relación (*causal/funcional*) con algún caso de una fórmula que expresa la proposición. Soy conciente de que hay toda una tradición filosófica en contra. Platón dice (creo) que hay una facultad intelectual especial (*theoría*) con la cual uno espía los objetos abstractos. Frege dice que uno *aprehende* (lo que estoy llamando) proposiciones, pero no puedo encontrar ninguna doctrina que diga en qué consiste la aprehensión, más allá del comentario (en "The Thought") según el cual no es percepción sensorial porque sus objetos son abstractos y no es introspección porque sus objetos no son mentales. (También dice que aprehender un pensamiento no es como agarrar un martillo. Seguramente.) En lo que a mí respecta, deseo un mecanismo para la relación entre organismos y proposiciones, y el único que se me ocurre es la mediación a través de representaciones internas.[14]

14. La idea de que la aprehensión de proposiciones es mediada por objetos lingüísticos no es enteramente atípica, ni siquiera para la tradición platónica. Dice Church: "La preferencia por (digamos) el ver sobre el *comprender* como método de observación, me parece caprichosa. Porque así como un cuerpo opaco puede ser visto, de la misma manera un concepto puede ser comprendido o aprehendido... En ambos casos, la observación no es directa sino mediante intermediarios... expresiones lingüísticas, en el caso del concepto." (1951); véase también la discusión en Dummett (1973, págs. 156-157).

Objeción 2: Seguramente es posible *concebir* que las actitudes proposicionales *no* sean relaciones con representaciones internas.

Pienso que lo es; la teoría de las actitudes proposicionales es una pieza de psicología empírica, no un análisis. Puesto que podría haber habido ángeles o el conductismo podría haber sido verdadero, y entonces la tesis de la representación interna habría sido falsa. La moraleja es, pienso, que debemos renunciar al análisis; la psicología es toda la filosofía de la mente con la que probablemente podamos contar.

Más aún, es *empíricamente* posible que haya criaturas que tengan las mismas actitudes proposicionales que nosotros (por ejemplo, las mismas creencias) pero que *no* tengan el mismo sistema de representaciones internas; criaturas que, por decirlo así, compartan nuestros estados cognoscitivos pero no nuestra psicología. Supóngase, por ejemplo, que los marcianos, o los delfines, resultan creer lo mismo que nosotros pero tienen un tipo de cálculo de costo muy distinto. Podríamos entonces inclinarnos a decir que hay relaciones de traducción entre los sistemas de representación interna (es decir, que representaciones formalmente distintas pueden expresar la misma proposición). Queda por ver si esa afirmación tiene o no sentido; es difícil que podamos pensar en ello sin antes elaborar teorías acerca de cómo tales sistemas son semánticamente interpretados; y, tal como están las cosas, no tenemos teorías semánticas de los lenguajes naturales, mucho menos de los lenguajes del pensamiento. Tal vez se presuponga que no constituye una objeción a una doctrina el que *pueda* llevar a incoherencias. O, mejor dicho, si es una objeción, existe una respuesta adecuada a ella: "Sí, pero también puede ser que no lo haga".

Terminaré con lo ya dicho. La psicología cognitiva contemporánea es, en efecto, un resurgimiento de la teoría representacional de la mente. El tratamiento preferido de las APs surge en este contexto. En particular, la mente es concebida como un órgano cuya función es la manipulación de representaciones, y éstas, a su vez, proveen el dominio de los procesos mentales y los objetos (inmediatos) de los estados mentales. Esto es lo que significa ver a la mente como algo parecido a un computador. (O, más bien, poniendo el caballo adelante del carro, esto es lo que significa ver a un computador como algo parecido a la mente. Damos sentido a la analogía al tratar a los estados seleccionados de la máquina como fórmulas, y al especificar las interpretaciones semánticas que reciben las fórmulas. Es en el contexto de tales especificaciones que hablamos de los procesos de la máquina como computaciones y de los estados de la máquina como intensionales [*intensional*].)

Si la teoría representacional de la mente es verdadera, entonces sabemos qué son las actitudes proposicionales. Pero el total neto de los problemas filosóficos no ha disminuido con ello. Ahora debemos enfrentar lo que siempre ha sido *el* problema para las teorías representacionales: ¿cómo se relacionan las representaciones internas con el mundo? ¿Qué significa para un sistema de representaciones internas estar semánticamente interpretado? Considero que este problema constituye hoy en día el tema principal de la filosofía de la mente.

TRADUCTORA: Eleonora Orlando.
REVISIÓN TÉCNICA: Eduardo Rabossi.

REFERENCIAS BIBLIOGRÁFICAS

Berlín, B., y P. Kay: (1969) *Basic Color Terms*. Berkeley y Los Angeles: University of California Press.

Brown, R.: (1976) "Reference — In Memorial Tribute to Eric Lenneberg", *Cognition* 4: 125-153.

Brown, R., y E. Lenneberg: (1954) "A study in Language and Cognition", *Journal of Abnormal and Social Psychology* 49, 454-462.

Carnap, R.: (1947) *Meaning and Necessity,* Chicago, University of Chicago Press, Phoenix Books.

Church, A.: (1951) "The Need for Abstract Entities in Semantic Analysis", en *Contribution to the Analysis and Synthesis of Knowledge. Proceedings of the American Academy of Arts and Sciences*, N° 80, págs. 100-112.

Davidson, D.: (1965) "Theories of Meaning and Learnable Languages", en Y. Bar Hillel (comp.): *Logic, Methodology and Philosophy of Science. Proceedings of the 1964 International Congress*, págs. 383-394, Amsterdam, North Holland.

Dennett, D.: (1969) *Content and Consciousness*, Londres, Routledge & Kegan Paul.

Dummett, M.: (1973) *Frege,* Londres, Duckworth.

Fodor, J. A.: (1975) *The Language of Thought,* Nueva York, Crowell.

Goodman, N.: (1968) *Languages of Art,* Indianapolis, Bobbs-Merrill.

Grice, H. P.: (1975) "Method in Philosophical Psychology". *Proceedings and Addresses of the American Philosophical Association*, vol. 48, págs. 23-53.

Heider, E.: (1972) "Universals in Color Naming and Memory", *Journal of Experimental Psychology* 93, 10-20.
Nagel, T.: (1965) "Physicalism", *Philosophical Review* 74, 339-356.
Vendler, Z.: (1972) *Res Cogitans,* Ithaca, Cornell University Press.

CAPÍTULO 7

LA TEORÍA SINTÁCTICA DE LA MENTE * (SELECCIÓN)

Stephen P. Stich

La teoría fuerte representacional de la mente

1. Las generalizaciones cognitivas expresadas en términos de contenido

[...] Ninguno de los autores que han defendido la Teoría Fuerte Representacional de la Mente [TRM Fuerte] [*Strong RTM*] ha sido más elocuente o explícito que Jerry Fodor.** Permítaseme comenzar la discusión de ese punto de vista reuniendo unas pocas citas tomadas de las obras de Fodor. (Como tendré oportunidad de volver sobre estas citas más adelante, las numeraré para facilitar la referencia.)

1

> Hemos sido conducidos al funcionalismo... por la sospecha de que existen generalizaciones empíricas acerca de los estados mentales que no pueden ser formuladas en el vocabulario de las teorías neurológicas o físicas... Pero cuando pensamos cómo son esas generalizaciones resulta sorprendente que todos los candidatos —literalmente *todos* ellos— son generalizaciones que valen para las actitudes proposicionales en virtud del contenido de las actitudes proposicionales. Para formular el punto, no necesitamos contar con ejemplos agudos tomados de la lingüística o de la psicología. Las etiologías psicológicas de sentido común [*commonsense psychological etiologies*] serán suficientes. Consideremos esto: ver que a es F es una causa normal para creer que

* Selección de los capítulos 7 y 8 de *From Folk Psychology to Cognitive Science*, Bradford Books, MIT Press. Con autorización del autor y de MIT Press.

** Adviértase que 'consistente' no se encuentra entre los adjetivos que he usado para caracterizar la prédica de Fodor, y con buenas razones, dado que Fodor puede ser (y ha sido) leído como abogando tanto en favor de la TRM Fuerte como de la TRM Débil.

a es F; ...los enunciados [que afirman] que P son causados normalmente por creencias de que P; ...etcétera, etcétera. Por cierto que el propósito de tales ejemplos no es [sostener] que, probablemente, figuren en una psicología cognitiva seria, sino más bien que nuestros intentos por lograr una psicología cognitiva seria se fundan en la esperanza de que *esa clase* de generalización puede ser sistematizada y formulada con rigor...Y: UNO NO PUEDE SALVAR TALES GENERALIZACIONES SIN APELAR A LA NOCIÓN DE CONTENIDO DE UN ESTADO MENTAL, dado que, como se señaló antes, esas generalizaciones son tales en tanto se aplican a los estados mentales en virtud de sus contenidos (Fodor, 1981b, págs. 25-26).

2

La situación paradigmática —la que resulta provechosa al cognitivista— es aquella en la cual las actitudes proposicionales interactúan causalmente *en virtud* de su contenido.
[...] Si hay contrafácticos contingentes y verdaderos que relacionan a los estados mentales-*caso* [*mental state tokens*] en virtud de sus contenidos, ello se debe presumiblemente a que hay generalizaciones contingentes y verdaderas que relacionan a los estados mentales-*tipo* [*mental state types*] en virtud de sus contenidos (Fodor, 1978, págs. 505 y 506).

3

Para esta discusión han tenido relevancia tres concepciones [...]: la idea de que los estados mentales son definidos funcionalmente; la idea de que para especificar las generalizaciones que las etiologías mentalistas instancian, es necesario referir [*advert*] a los contenidos de los estados mentales, y la idea de que los estados mentales son relaciones con representaciones mentales, entendidas estas últimas, *inter alia*, como objetos semánticamente interpretados (Fodor, 1981b, pág. 30).

En estos pasajes hay dos ideas que adquirirán una importancia considerable en la discusión que sigue. La primera es que la "psicología cognitiva seria" está fundada en la esperanza de que las generalizaciones empíricas de la psicología de sentido común [*commonsense psychology*] puedan ser sistematizadas y formuladas con rigor. La segunda es que las generalizaciones de la psicología de sentido común y también las gene-

ralizaciones de la ciencia cognitiva "referirán a los contenidos de los estados mentales". Esta última idea es esencial para lo que he estado denominando TRM Fuerte. La afirmación de que la ciencia cognitiva busca (o debería buscar) "generalizaciones que relacionen estados mentales-tipo en virtud de sus contenidos", es lo que hace *fuerte* a la TRM Fuerte.

[...] Diré algo acerca de qué es lo que hace *representacional* a la TRM Fuerte. La idea básica es que los estados mentales —tanto los postulados por la psicología de sentido común [*folk psychology*] como los postulados por la ciencia cognitiva— han de ser vistos como relaciones con alguna especie de entidades representacionales. Las oraciones en un código mental o en un lenguaje del pensamiento, son los candidatos obvios para el papel de representaciones mentales. Pero la TRM Fuerte no precisa enfatizar que las representaciones sean oraciones. Otras clases de representaciones podrían también servir, en tanto sean la clase de cosas de las que puede pensarse que tienen contenido o que son semánticamente interpretadas. Si las entidades representacionales son oraciones entonces los estados mentales (tanto los de sentido común como los postulados por la ciencia cognitiva) han de ser considerados como relaciones entre un organismo y un caso de una oración... Ahora quiero considerar la afirmación de que la ciencia cognitiva busca generalizaciones que especifiquen relaciones causales entre los estados mentales en términos de sus contenidos.

Teniendo presente la diatriba formulada en el capítulo 4, no causará sorpresa al lector saber que me siento un tanto incómodo con la mención que hace Fodor a las generalizaciones que "refieren a los contenidos" de los estados mentales y a los estados mentales que interactúan causalmente "en virtud de su contenido". Porque esto suena fuertemente como si los contenidos fueran cosas —un tipo de entidad— sobre las que las teorías cognitivas pueden cuantificar. Pero creo que en este caso las apariencias son engañosas. Fodor ni explica ni usa seriamente la idea de los contenidos *en tanto entidades*, y creo que su discurso acerca de las generalizaciones que relacionan estados "en virtud de sus contenidos" se entiende mejor como una abreviatura de generalizaciones que relacionan estados *en virtud de las oraciones-contenido* [*content sentences*] *que usamos para caracterizarlos*...

2. Algunas razones para sospechar de la TRM Fuerte

El resto de este capítulo estará dedicado a exponer las razones por las que debería rechazarse el modelo de ciencia cognitiva que propone

la TRM Fuerte. Lo que trataré de mostrar es que si insistimos en expresar nuestras generalizaciones cognitivas en términos de contenido, perderemos generalizaciones significativas y poderosas, y nos veremos enfrentados a una vaguedad endémica y a menudo dañosa. Ninguno de los argumentos que siguen prueba que no podría ser *posible* construir una ciencia cognitiva siguiendo las líneas propuestas por la TRM Fuerte. No se me ocurre cómo podría argüirse en favor de una conclusión tan vasta. En cambio, espero convencer al lector de que el científico cognitivo tiene que pagar un precio muy alto por adherir a la TRM Fuerte...

Antes de exponer mi caso detalladamente puede ser útil poner de manifiesto algunas de la razones que a primera vista llevan a sospechar acerca de la utilidad científica de las generalizaciones expresadas en términos de contenido. Creo que tales razones están justificadas. En primer lugar... en las adscripciones de contenido propias del sentido común, está involucrada una apelación a la similitud [*similarity*]. En consecuencia, los predicados de la forma 'cree que p' son a la vez *vagos* y *sensibles al contexto* [*context sensitive*], como lo son los predicados tales como 'se parece a Abraham Lincoln'. Así, habrá casos en los que el predicado 'cree que p' se aplica con claridad, pero también habrá muchos casos en los que, fuera de contexto, no hay manera de decir si el predicado se aplica o no. De tal modo, si las generalizaciones de una teoría cognitiva son moldeadas en términos de tales predicados, a menudo resultará poco claro si las generalizaciones se aplican o no a un sujeto determinado.

Hay una segunda manera en la cual 'cree que p' se asemeja a predicados como 'se parece a Abraham Lincoln'. En ambos casos se efectúa una comparación con un estándar o con un ejemplar. En el caso de creer, el estándar somos *nosotros mismos*. Creer que p es estar en un estado creencial [*belief state*] similar al que subyace a nuestra aserción sincera de 'p'. De tal modo, hay una suerte de *relatividad respecto del observador* [*observer relativity*] inserta en nuestra noción de sentido común, y la teoría cognitiva escrita en el lenguaje de creencia de sentido común heredará tal relatividad. Esto tiene dos desventajas. Lo primero y más obvio es el hecho de que distintos observadores pueden diferir substancialmente entre sí, y que cuando ello acaece pueden ser llevados a describir de maneras diferentes a las creencias del sujeto. O lo que quizás es más fácil de ilustrar, puede haber casos en los que un observador no disponga, sencillamente, de ninguna oración-contenido tal que la pueda usar con comodidad para caracterizar las creencias del sujeto. Esta última observación sugiere otra dificultad, que considero más seria

aún. Dado que el lenguaje de creencia de sentido común caracteriza al estado cognitivo de un sujeto comparándolo con el nuestro, aquellos sujetos que difieran de nosotros de manera muy extrema quedarán totalmente fuera del alcance de la descripción [en términos] de sentido común. Así, si nuestra teoría cognitiva descansa esencialmente en el vocabulario de la psicología de sentido común que adscribe contenido, esos sujetos quedarán más allá de su alcance. Sin embargo, algunos de esos sujetos pueden tener mentes que funcionen de manera muy parecida a la nuestra. Si hay generalizaciones importantes que abarcan a nuestras personas y a esos seres exóticos, una ciencia cognitiva moldeada a la manera de la TRM Fuerte está destinada a omitirlos. Y si hay generalizaciones específicas para una o para otra categoría de sujetos tales, ellas serán difíciles o imposibles de enunciar de la manera requerida por la TRM Fuerte.

Otra razón prima facie para tener sospechas acerca de la utilidad científica de las generalizaciones que refieren al contenido versa sobre el papel de la similitud ideológica y de la referencia en la individuación de las creencias. A menudo, ambos factores imponen un esquema individuador de grano mucho más fino [*fine-grained*] del que se requeriría si usáramos nociones individuadas de acuerdo con lineamientos causales estrechos [*narrow causal lines*]. Las distinciones impuestas poseen una utilidad incuestionable si lo que nos interesa son las propiedades semánticas de las creencias y de las oraciones que las expresan. Pero si nuestro interés reside en predecir y explicar la conducta, hay motivos para sospechar que esas distinciones, sencillamente, se interpondrán en el camino. Nos forzarán a atribuir rótulos diferentes a estados psicológicos que son idénticos entre sí en lo que hace a potencial causal [*causal potential*]. De tal modo, será más difícil capturar generalizaciones conspicuas [*salient*] para la explicación de la conducta. En vez de luchar con esas distinciones de grano fino que en nada contribuyen a la explicación de la conducta, un teórico cognitivo podría verse tentado a amputarlas y a construir una teoría con el concepto más austero que reste. Por cierto, ... eso es lo que se hace de manera corriente en la práctica efectiva de los teóricos cognitivos. Pero, por supuesto, un teórico que al enunciar las generalizaciones abandone los componentes ideológicos y referenciales del contenido, no se encontrará ya trabajando dentro del paradigma explicativo propuesto por la TRM Fuerte...

La teoría sintáctica de la mente [*TSM*]

1. El enfoque TSM de las teorías cognitivas

La idea básica de la TSM es que los estados cognitivos cuya interacción es (en parte) responsable de la conducta, pueden ser mapeados [*mapped*] sistemáticamente en objetos abstractos sintácticos de manera tal que las interacciones causales entre los estados cognitivos, así como los nexos causales con los estímulos y los eventos conductuales, pueden ser descriptos en términos de las propiedades y relaciones sintácticas de los objetos abstractos en los que los estados cognitivos son mapeados. En pocas palabras, la idea es que las relaciones causales entre los estados cognitivos reflejan [*mirror*] relaciones formales entre objetos sintácticos. Si esto es correcto, será natural considerar a los estados cognitivos-caso [*cognitive state tokens*] como casos de objetos abstractos sintácticos.

Todo esto requiere ser expuesto de modo mucho más cuidadoso y para hacerlo tendré que decir bastante acerca de los estados. De tal manera, es mejor que comience diciendo cómo debería interpretarse este discurso. Tal como yo los concibo, los estados son la instanciación [*instantiation*] de una ***propiedad*** por un *objeto* durante un ***intervalo de tiempo***. Así interpretados, los estados son ítemes *particulares* [*particulars*] con una ubicación más o menos definida en el espacio y en el tiempo. De acuerdo con el enfoque que adopto, los estados son susceptibles de lo que podría denominarse una clasificación *esencial* en tipos. Un par de estados son del mismo *tipo esencial* [*essential type*] si y sólo si son instanciaciones de la misma propiedad. A veces será conveniente usar la palabra 'estado' para denotar la propiedad que tienen en común todos los estados del mismo tipo esencial. De acuerdo con esta manera de hablar, cuando digo que un cierto organismo está *en el estado P*, significaré que el organismo instancia la propiedad P en el momento en cuestión. De modo similar, decir que un par de organismos están *en el mismo estado* equivale a decir que instancian la misma propiedad esencial en el momento en cuestión. Cuando la ambigüedad nos amenace resultará útil usar 'estado-caso' [*state token*] para hacer referencia a los ítemes particulares y 'estado-tipo' [*state type*] para hacer referencia a las propiedades (véase Kim, 1969 y 1976).

Aunque cada estado-caso tiene sólo un único tipo esencial, los estados, como otros ítemes individuales, pueden ser agrupados de maneras infinitamente variadas en tipos no esenciales. El tipo de un estado-caso es, simplemente, una categoría de ítemes individuales, y habremos espe-

cificado tal tipo cuando hayamos establecido las condiciones para ser un miembro de la categoría. De manera similar, los estados-tipo [*state types*], tanto esenciales como no esenciales, pueden ser agrupados en tipos o categorías. Para especificar una categoría de estados-tipo se requiere especificar la propiedad que un tipo tiene que tener para contar como miembro de la categoría. Una última dificultad es que el tipo o categoría de los estados-tipo impone una categorización derivativa o indirecta en los estados-caso: si ϕ es una propiedad cuya posesión por parte de un estado-tipo es necesaria y suficiente para que el estado-tipo sea de la categoría C, entonces podríamos pensar a todos los casos de los estados-tipo en C, como casos de ϕ. *Ser un caso de* ϕ, entonces, es la propiedad que tiene un estado-caso en virtud de ser de un tipo que tiene en sí mismo una determinada propiedad. (Adviértase que una categorización de naturaleza derivativa o indirecta análoga es moneda corriente cuando se habla acerca de tipos y de casos lingüísticos. La palabra (tipo) 'lingüístico' cae dentro de la categoría de los adjetivos. Y así, el caso de 'lingüístico' que aparece en la oración anterior, puede pensarse también como un caso de un adjetivo.) Esto completa mi breve excursión por la ontología. Volvamos ahora a la TSM.

Puede considerarse que la tarea que tiene el teórico al establecer una teoría cognitiva TSM, consta de tres partes. Primero, tiene que especificar una clase de objetos sintácticos (tipos, por supuesto, no casos) de modo tal que asigne una estructura formal o sintáctica a cada uno de dichos objetos. Dado que corrientemente habrá en la clase un número infinito de objetos, eso se hace mejor con una gramática o un conjunto de reglas de formación que detallen las maneras según las cuales se pueden construir objetos sintácticos complejos a partir de un conjunto finito de primitivos.

Segundo, el teórico hipotetiza que para cada organismo abarcado por la teoría existe un conjunto de estados-tipo cuyos casos están causalmente implicados en la producción de la conducta. También hipotetiza que hay un mapeo de esos estados-tipo en los objetos sintácticos de la clase especificada. Corresponde formular varias observaciones acerca de estas hipótesis. En primer lugar, el teórico necesita decir muy poco acerca de la naturaleza esencial de los casos de los estados que están causalmente implicados en la producción de la conducta. Presumiblemente son estados físicos del cerebro y así las propiedades que constituyen sus tipos esenciales son propiedades neurológicas; aunque un teórico TSM que desee ser recatado o cauteloso no necesita siquiera comprometerse con ese punto. (De aquí en más supondré que los esta-

dos causalmente implicados en la producción de la conducta son estados neurológicos, pues ello simplificará considerablemente mi exposición.) En segundo lugar, cuando se afirma la existencia del mapeo, el orden de los cuantificadores posee alguna importancia. El teórico no pretende que el mapeo sea el mismo para cada sujeto, sino sólo que para cada sujeto haya un mapeo. Así, en sujetos diferentes, estados-tipo neurológicos muy diferentes pueden ser mapeados en un objeto sintáctico dado. Estas primeras dos observaciones expresan, por supuesto, el espíritu del funcionalismo, que enfatiza la posibilidad de las realizaciones múltiples [*multiple realizations*] de los estados mentales. Una tercera observación es que no es necesario que el teórico se limite a alegar que hay un mapeo único de los estados neurológicos en su clase preferida de objetos formales. En cambio, el teórico puede afirmar que hay diversas categorías de estados neurológicos y que los estados en cada categoría son mapeados en los objetos formales. Por ejemplo, si la teoría se inspira en la psicología de sentido común, puede postular dos clases [*classes*] distintas de estados que subyacen a la conducta, estados similares a creencia [*belief-like states*] y estados similares a deseo [*desire-like states*], tal que los estados-tipo de *ambas* clases sean mapeados en una única clase de objetos formales. Así, un estado-tipo dado similar a creencia y un estado-tipo dado similar a deseo pueden ambos ser mapeados en el mismo objeto formal. Finalmente, se supone que el mapeo a partir de la categoría similar a creencia y la categoría similar a deseo, abarca todos los posibles estados-tipo en esa categoría, y se supondrá en general que en cada categoría son posibles un número infinito de estados-tipo distintos.

La tercera parte de una teoría cognitiva construida de acuerdo con las pautas de la TSM, es la especificación de las generalizaciones de la teoría. La idea central de la TSM —la idea que la hace *sintáctica*— es que las generalizaciones que detallan las relaciones causales entre los estados neurológicos hipotetizados han de ser especificadas de manera indirecta *via* las relaciones formales entre los objetos sintácticos en los que se mapean los estados-tipo neurológicos. De manera similar, las generalizaciones que especifican las relaciones causales entre los estímulos y los estados neurológicos no identificarán a los estados neurológicos por referencia a sus tipos neurológicos esenciales sino, en cambio, por referencia a los objetos sintácticos en los que se mapean los tipos de estado neurológicos. Lo mismo vale para las generalizaciones que especifican las relaciones causales entre los estados neurológicos y la conducta.

Es posible que en este punto el lector experimente una sensación de *déjà vu* dado que si los objetos sintácticos elegidos por el teórico son *oraciones* entonces los estados neurológicos postulados por una teoría a la manera de la TSM cuadran con las *oraciones mentales individuadas causalmente de manera estrecha* [*narrow causally individuated mental sentences*]... Dado que los estados-tipo neurológicos hipotetizados son mapeados en oraciones-tipo [*sentence types*], los casos de esos estados neurológicos podrían ser considerados, plausiblemente, como casos (en el sentido derivativo) de las oraciones-tipo con las que es apareado su tipo neurológico. Dado que puede haber categorías diferentes de estados neurológicos hipotéticos, un sujeto puede tener entre sus estados mentales más de un caso de un tipo dado de oración. Puede, por ejemplo, tener un caso de una oración— tipo que es un estado similar a creencia y otro que es un estado similar a deseo; o puede tener un estado similar a deseo que es un caso de una oración condicional y un estado similar a creencia que es un caso del antecedente del condicional.

[...] Dado que la motivación para concebir a los estados-caso neurológicos hipotéticos como oraciones-caso es describir las relaciones causales haciendo referencia a casos sintácticos, tenemos que preguntar *qué* relaciones sintácticas tienen que ser reflejadas por los estados-caso neurológicos para que valgan como oraciones-caso. Existen, creo, tres respuestas muy diferentes que podrían darse a tal pregunta. Una idea consiste en insistir que si un estado-caso neurológico ha de valer como caso de una oración, tiene que satisfacer *todas* las generalizaciones especificadas por la teoría. Esta estrategia tiene una desventaja considerable dado que aún los cambios pequeños en las generalizaciones de la teoría implicarán una modificación en la especificación de qué es ser un caso de una oración mental-tipo. Otra idea, que evita esta dificultad, es especificar un conjunto de generalizaciones *esenciales* que un estado neurológico tiene que satisfacer si sus casos han de valer como casos de una oración-tipo dada. Puede agregarse y modificarse generalizaciones adicionales, tanto como sea necesario, sin alterar la especificación relativa a tipos [*the account of typing*]. Pero este enfoque también tiene sus limitaciones. Es difícil visualizar qué motivación puede haber para distinguir un conjunto especial de generalizaciones como esenciales y es difícil visualizar cómo la separación entre generalizaciones esenciales y no esenciales podría dejar de ser arbitraria. Una tercera idea consiste en evadir el tema enfatizando solamente que para valer como un caso de una oración-tipo, un estado neurológico tiene que satisfacer a un número substancial de las generalizaciones incluidas en una teoría, sin espe-

cificar las generalizaciones particulares que tienen que ser satisfechas y sin estipular con exactitud cuántas de ellas tienen que serlo. Esto evita el problema que afecta a la primera estrategia porque permite modificaciones en el conjunto de las generalizaciones sin cambiar nuestro relato acerca de qué es lo que hace que un estado-caso valga como una oración-caso. También evita la arbitrariedad de la segunda estrategia, pero lo hace al costo de introducir un elemento de vaguedad en la especificación relativa a tipos. Los teóricos que desarrollan en la práctica los lineamientos adelantados por la TSM tienden a preocuparse muy poco por el problema, aunque sospecho que si se los presionara optarían por la tercera estrategia como la que mejor representa sus intenciones.

Debe advertirse que hay una especie de holismo [*holism*] involucrado en las tres estrategias destinadas a especificar tipos [*typing*] para casos mentales. Es sólo contra el trasfondo de un mapeo sistemático de estados-tipo a oraciones-tipo que un estado-caso dado vale como el caso de una oración-tipo específica. O para decirlo de otra manera: ningún estado neurológico puede valer como un caso de una oración-tipo, a menos que muchos estados neurológicos valgan como casos de muchos tipos de oraciones-tipo diferentes...

2. Las ventajas de las teorías TSM

Lo que alego es que las teorías TSM son un candidato mejor para el teórico cognitivista que las teorías cuyas generalizaciones apelan al contenido, dado que las teorías sintácticas pueden dar cuenta de todas las generalizaciones abarcables al cuantificar sobre las oraciones-contenido, evitando, al mismo tiempo, las limitaciones que impone el lenguaje de contenido de sentido común. Así, las teorías TSM pueden abarcar generalizaciones que caen fuera del alcance de las teorías moldeadas según la TRM Fuerte... Para decirlo en términos simples, *la virtud de las teorías TSM consiste en que eliminan al intermediario.* Los estados mentales postulados por la teoría TSM no son caracterizados por sus oraciones-contenido sino, en cambio, por los objetos sintácticos en los que son mapeados. Aquellos pueden ser seleccionados por el teórico con miras a dar una explicación más simple y más poderosa de los nexos causales entre los estímulos, los estados mentales y las conductas, sin preocuparse por las similitudes o disimilitudes entre el sujeto y el teórico. Al eliminar al intermediario las teorías TSM pueden caracterizar a los estados cognitivos de un sujeto en términos apropiados al propio sujeto en vez de caracterizarlos en términos que fuercen una compara-

ción entre el sujeto y nosotros mismos. Y esto elimina el problema central de las teorías TRM Fuerte, dado que no hay peligro de que se pierdan generalizaciones cuando los sujetos sean tan diferentes de nosotros que la psicología de sentido común no pueda describirlos. Más aún, al eliminar la apelación a variadas dimensiones de *similitud* también se elimina gran parte de la vaguedad que afecta a las teorías cognitivas basadas en el contenido [*content-based cognitive theories*]...

3. El Solipsismo Metodológico y el Principio de Autonomía

Permítaseme resumir el punto al que nos ha conducido la argumentación de la parte II. La cuestión crucial es si la noción de creencia y otras nociones psicológicas de sentido común encontrarán una ubicación confortable en la ciencia cognitiva. Un punto de vista que propone una respuesta afirmativa es la Teoría Fuerte Representacional de la Mente, que visualiza a la ciencia cognitiva madura como postulando estados representacionales y refiriendo en sus generalizaciones al contenido. Sin embargo, en el capítulo 7 hemos reunido un conjunto de argumentos que se proponen mostrar que no es aconsejable que el científico cognitivo adopte el paradigma de la TRM Fuerte. El costo que se paga en términos de vaguedad y de pérdida de generalizaciones es muy elevado. En este capítulo he argumentado que existe una alternativa mejor. La Teoría Sintáctica de la Mente, al no apelar al contenido en las generalizaciones cognitivas, soslaya las dificultades que afectan a la TRM Fuerte. En esta sección quiero fortalecer la defensa de la TSM como alternativa a la TRM Fuerte, ofreciendo un par de argumentos. Cada uno de estos argumentos defiende un principio acerca de cómo deberían ser las teorías psicológicas. Los principios, *solipsismo metodológico* y *principio de autonomía*, están íntimamente relacionados y cada uno de ellos implica claramente que la psicología cognitiva no debe aspirar a expresar sus generalizaciones en términos de contenido. Sin embargo, los argumentos en favor de los principios son muy diferentes en cuanto a estrategia y plausibilidad. Aunque ninguno de ellos pretenda ser apodíctico, me inclino a pensar que el argumento a desarrollar en favor del principio de autonomía es significativamente más persuasivo que el argumento en favor del solipsismo metodológico. Quizás esto se deba a que el primer argumento es mío y el segundo es de Jerry Fodor. Lo incluyo aquí no porque piense que agrega mucho peso a la defensa de la TSM frente a la TRM Fuerte, sino porque ha sido discutido tan extensamente que no

puede ser ignorado en estas materias. Comenzaré con el solipsismo metodológico.

La expresión 'solipsismo metodológico' fue introducida originariamente por Putnam (1975) para caracterizar un punto de vista que deseaba criticar. De acuerdo con el planteo de Putnam, hay una distinción entre los "estados psicológicos en el sentido amplio" ["*the wide sense*"] y los "estados psicológicos en el sentido estrecho" ["*the narrow sense*"] (pág. 137). Los estados psicológicos en el sentido estrecho no presuponen "la existencia de ningún individuo, salvo la del sujeto al que el estado es adscripto" (pág. 136). Los estados psicológicos en el sentido amplio presuponen la existencia de algún otro objeto o individuo. El dolor podría ser un ejemplo natural de estado psicológico estrecho, mientras que tener celos de Enrique es prima facie un ejemplo de un estado psicológico en el sentido amplio dado que implica la existencia de Enrique. (Por supuesto, en términos estrictos no es el estado psicológico el que implica la existencia de Enrique; más bien, el enunciado de que el estado psicológico se da, implica el enunciado de que Enrique existe.) La doctrina del solipsismo metodológico sostiene que la psicología debe ocuparse exclusivamente de los estados psicológicos en el sentido estrecho. El peso del argumento de Putnam es que el solipsismo metodológico es insostenible pues excluye de la psicología estados tales como conocer el significado de un término.

Fodor, por el contrario, nos urge a que adoptemos al solipsismo metodológico como una estrategia de investigación en la psicología cognitiva. Pero en las manos de Fodor la noción de solipsismo metodológico es objeto de una elaboración importante. Su tesis central es que los estados y procesos mentales, o al menos aquellos a los que la psicología cognitiva debe atender, son "computacionales" (1980a, pág. 226). "Los procesos computacionales son, a la vez, *simbólicos* y *formales*. Son simbólicos porque son definidos sobre representaciones, y son formales porque se aplican a las representaciones en virtud (aproximadamente) de su *sintaxis*" (pág. 226). "Más aún, lo que hace de las operaciones sintácticas una especie de las operaciones formales es que ser sintácticas es una manera de *no* ser semánticas. Las operaciones formales son aquellas que son especificadas sin hacer referencia a propiedades semánticas de las representaciones tales como verdad, referencia y significado" (pág. 227). Finalmente, pareciera que para Fodor el solipsismo metodológico es simplemente la doctrina de que *la psicología cognitiva debe restringirse a postular operaciones formales sobre los estados mentales*. Ella no debe postular procesos que valgan para los estados men-

tales en virtud de sus propiedades semánticas. Fodor concede con franqueza que la doctrina del solipsismo metodológico es poco precisa dado que él no puede proporcionar ni un criterio ni una enumeración completa de qué va a valer como una propiedad semántica.

Debería resultar claro que el solipsismo metodológico congenia a la perfección con la Teoría Sintáctica de la Mente. También implica el rechazo de la TRM Fuerte. Porque en toda explicación plausible de qué va a valer como semántico, el teórico que expresa sus generalizaciones (su explicación de los procesos mentales) en términos de las *oraciones-contenido* utilizadas para caracterizar los estados mentales, postula operaciones mentales cuya especificación requiere que se haga referencia a las propiedades semánticas de dichos estados. El caso más claro es, supongo, el de la referencia: la noción semántica por excelencia. Dado que la similitud de la referencia es uno de los rasgos que determinan la corrección [*propriety*] de una oración-contenido, toda operación mental cuya especificación torne apropiadas a las oraciones-contenido respecto de los estados involucrados, entrará en colisión con los escrúpulos del solipsista metodológico. Por el contrario, las teorías cognitivas moldeadas según la TSM son casos paradigmáticos del tipo de teoría que el solipsista metodológico aprobaría. Parecería, pues, que una argumentación adecuada en favor del solipsismo metodológico nos tendría que proveer de una razón adicional para preferir la TSM a la TRM Fuerte.

Para nuestros fines, la parte esencial del argumento de Fodor es su defensa de la condición de formalidad, que requiere que las propiedades semánticas de los estados mentales no desempeñen ningún rol en la especificación de las generalizaciones psicológicas. Desafortunadamente, lo que Fodor dice sobre este tema es muy poco agudo. Tal como yo lo entiendo, Fodor argumenta así. Primero, si un estado mental tiene propiedades semánticas, ellas están fijadas presumiblemente por una o más "relaciones orgánico/ambientales" ["*organism/environment relations*"] (pág. 244). Segundo, los psicólogos que se mofan de la condición de formalidad y rechazan el solipsismo metodológico (Fodor los llama "naturalistas") "se proponen hacer ciencia a partir de relaciones orgánico/ambientales que (presumiblemente) fijan propiedades semánticas" (pág. 244). Tercero, para hacer esto el naturalista "tendría que definir las generalizaciones sobre los estados mentales, de un lado, y sobre entidades ambientales, del otro" (pág. 249). Pero, cuarto, para definir tales generalizaciones el naturalista tiene que tener alguna "manera canónica de referirse a las segundas" y así, cuando las entidades ambientales fueran descriptas de tal manera, tendría que hacer que las generalizaciones

fueran "instanciadoras de leyes" [*law-instantiating*] (pág. 249). Dicho de otra manera, las caracterizaciones de los objetos del lado ambiental de la interacción orgánico/ambiental, tienen que ser caracterizaciones "proyectables" (pág. 250), que "expresen propiedades nomológicamente necesarias" (pág. 249) de los objetos. Sin embargo, en opinión de Fodor, esto último es lo que falla. Porque, quinto, para obtener tales caracterizaciones proyectables o instanciadoras de leyes, tenemos que esperar el desarrollo adecuado de la ciencia que estudia el objeto. Si el objeto es la sal, entonces la caracterización proyectable adecuada, esto es, 'ClNa', "sólo estará disponible *después* de que hayamos hecho química" (pág. 249). Pero como los objetos del ambiente podrían ser todo lo que podemos pensar o referir, no contaremos con caracterizaciones adecuadas de tales objetos hasta que todas las ciencias no psicológicas hayan terminado con su trabajo. "La teoría que caracteriza a los objetos del pensamiento es la teoría de *todo*; es la totalidad de la ciencia. En consecuencia... los psicólogos naturalistas heredarán la Tierra, pero sólo después de que todos los demás hayan terminado con ella" (pág. 248). No debemos intentar una psicología naturalista, concluye Fodor, porque tal tentativa tiene que esperar a que todas las demás ciencias proporcionen caracterizaciones proyectables de los objetos del ambiente que interactúan con el organismo. "Sin duda, es correcto tener una estrategia de investigación que diga 'espera algún tiempo'. Pero, ¿quién quiere esperar *por siempre*?" (pág. 248).

Aunque apoyaría con entusiasmo el principio del solipsismo metodológico, tengo reservas respecto del argumento que lo apoya. Si he reconstruido con corrección el argumento de Fodor, sospecho que hay dos lugares en los que es susceptible de crítica. El primero es el paso tercero, que sostiene que "hacer ciencia" de las relaciones orgánico/ambientales que determinan la referencia equivale a buscar *generalizaciones nomológicas* que enlazan entidades ambientales y estados mentales. Ésta es, sin duda, *una* manera en la que puede proceder una ciencia preocupada por las relaciones de fijación de la referencia; y como Fodor señala, eso es lo que han intentado hacer típicamente los psicólogos de talante naturalista (1980b, pág. 102). Pero, por lo que sé, no es necesario que quienes practican la ciencia de las interacciones orgánico/ambientales que subyacen a la referencia, busquen leyes causales. Después de todo, existen muchos dominios científicos respetables, desde la botánica descriptiva, la etología y la paleobiología a la antropología y la lingüística, en los que la búsqueda de generalizaciones nomológicas desempeña un rol relativamente menor. Volviendo a nues-

tro caso, es verdad, tal como sugiere el segundo paso del argumento de Fodor, que si un psicólogo expresa sus generalizaciones cognitivas en términos de oraciones-contenido apropiadas para distintos estados, entonces su teoría *involucrará*, de una u otra manera, las relaciones orgánico/ambientales que contribuirán a determinar la corrección de las oraciones-contenido. Sin necesidad de argumentación adicional: no es claro que el psicólogo que practica la estrategia de la TRM Fuerte tenga que buscar *generalizaciones nomológicas acerca de* las relaciones orgánico/ambientales que determinan la referencia y el contenido.

El segundo punto en el que temo que el argumento de Fodor sea vulnerable es el paso quinto, que afirma que las caracterizaciones proyectables apropiadas de los objetos del lado ambiental de las interacciones orgánico/ambientales sólo serán proveídas por las ciencias que estudian esos objetos. Fodor sugiere fuertemente, que la física y la química son las ciencias apropiadas para indagar acerca de clases naturales como el agua y la sal, y expresa extrañeza acerca de "qué ciencia [trataría] de tíos, paraguas y empresarios" (pág. 103). Quizá Fodor tenga más sensibilidad que yo para estas delicadas cuestiones, pero sin argumentación adicional no estoy convencido de que predicados de sentido común tales como 'sal', 'tío' y 'empresario' no sean candidatos respetables para ser incorporados en las generalizaciones nomológicas. Sin duda que no son los predicados proyectables ideales a ser usados en las generalizaciones de la física y de la química. Pero quienes, como nosotros, tomamos en serio a las ciencias especiales, tenemos la expectativa de que los esquemas clasificatorios invocados en esas ciencias se crucen con los núcleos clasificatorios impuestos por la física. Presumiblemente, la física y la química no van a tener generalizaciones que invoquen 'tío' o 'paraguas', pero la antropología y la economía podrían muy bien hallar mucho más útiles a esos términos. Irónicamente, Fodor ha sido un elocuente defensor de la respetabilidad científica de los esquemas clasificatorios que no se reducen fácilmente a los de las ciencias físicas (véase Fodor, 1974).

Pese a todo esto, soy un tanto renuente a rechazar sin más el argumento de Fodor a favor del solipsismo metodológico. Quizás el argumento pueda ser elaborado de manera tal que se ponga en claro por qué el psicólogo cognitivo que usa nociones psicológicas de sentido común tiene que buscar el tipo de generalizaciones requeridas por el paso tercero. Y quizás algo más pueda ser dicho para establecer que predicados cotidianos como 'sal' y 'paraguas' no son adecuados para formular generalizaciones orgánico/ambientales. Pero dada la ausencia

de tales elaboraciones me inclino a no otorgar al argumento demasiado peso.

Pasaré ahora al principio de autonomía. La idea básica del principio es que los estados y procesos que deben concernir al psicólogo son los que supervienen [*supervene*] al estado físico, interno, corriente [*current*] del organismo. (En términos generales, una clase de estados y procesos superviene a otra cuando la presencia o ausencia de los estados y procesos de la primera clase está totalmente determinada por la presencia o ausencia de los estados y procesos de la segunda. Para un análisis útil de la superveniencia, véase Kim, 1978 y 1982.) Esto equivale a afirmar que una teoría psicológica debe ignorar todas las diferencias entre organismos, tal que ellas mismas no se manifiesten como diferencias en [sus] estados físicos internos, corrientes. Si respetamos el principio de autonomía entonces el hecho de que un par de organismos tenga diferentes historias o de que estén en ambientes significativamente diferentes, será irrelevante para una teoría psicológica, salvo que esas diferencias sean relevantes para el estado físico interno, corriente, del organismo. O, para plantear la cuestión desde el punto de vista opuesto, los hechos históricos y ambientales serán psicológicamente relevantes sólo cuando influyan en el estado físico, interno, corriente, de un organismo. De modo que si un rasgo de la historia o del ambiente de un organismo podría haber sido diferente sin afectar el estado físico interno, corriente, del organismo, entonces ese rasgo histórico o ambiental no tendría que desempeñan ningún rol en la teoría psicológica.

El principio de autonomía es incompatible, lo mismo que el solipsismo metodológico, con la estrategia explicativa que propone la TRM Fuerte. El principio de autonomía prohíbe las generalizaciones expresadas en términos de las oraciones-contenido que caracterizan a los estados mentales, dado que la corrección de una oración-contenido en tanto caracterización de un estado mental, está determinada en parte por la similitud de la referencia. A su vez, la referencia está determinada, en parte, por las historias causales distantes [*distant*] de un término o concepto y, en parte, por el ambiente sociolingüístico en el que el sujeto está involucrado. Pero ninguno de esos factores necesita dejar su trazo en el estado físico interno, corriente, del organismo. Así, es posible que un par de sujetos difiera en la referencia de algún término que usan, aun cuando en sus estados físicos, internos, corrientes, no exista una diferencia que le corresponda. La condición de formalidad que promueve el solipsista metodológico prohíbe, directamente, generalizaciones que descansen en las propiedades semánticas de los estados a los que se apli-

can. El principio de autonomía psicológica logra en gran medida la misma meta al impedir la apelación a los factores históricos y ambientales de los que dependen en parte las propiedades semánticas, tales como la referencia.

En el trabajo en el que propuse el principio de autonomía (Stich, 1978) no ofrecí ningún argumento en su defensa pues me pareció que poseía una plausibilidad intuitiva substancial. Pero las discusiones subsiguientes pusieron en claro que la atracción intuitiva del principio de autonomía comienza a desvanecerse cuando la gente ve qué es lo que implica respecto del uso de las nociones psicológicas de sentido común en la psicología científica. De tal modo, que se requiere algún argumento. Pienso que la mejor defensa del principio de autonomía comienza con lo que podría denominarse el *argumento del reemplazo* [*replacement argument*]. Supongamos que tuviéramos éxito en construir una réplica exacta de mí: un cuerpo humano viviente cuyos estados físicos internos corrientes fueran idénticos, en cierto momento, a los míos en ese momento. Supóngase, además, que mientras que estoy profundamente dormido soy secuestrado y reemplazado por mi réplica. Parecería que si el delito fuera adecuadamente ocultado, nadie (con excepción de los raptores y de mí mismo) sabría nada. Porque la réplica, al ser una copia física exacta, se comportaría como lo haría yo en todos los casos. Ni siquiera la réplica sospecharía que es un impostor. Pero, el argumento continúa, dado que la psicología es la ciencia que aspira a explicar la conducta, los estados, procesos o propiedades que no son compartidos por Stich y por su réplica de idéntica conducta, tienen, sin duda, que ser irrelevantes para la psicología.

Creo que en el argumento del reemplazo hay un núcleo importante de verdad. Pero tal como se lo propone, no funciona. El problema es que en muchas circunstancias mi réplica y yo no nos comportamos (no podríamos comportarnos) de la misma manera; al menos no lo haremos tal como nuestra conducta sería descripta de manera corriente. Un ejemplo servirá para aclarar el punto. Uno de mis bienes es un viejo automóvil. Si usted me ofreciera mil dólares por él me encantaría vendérselo de inmediato. Pero supongamos que he sido secuestrado antes de la oferta y que mi réplica ha ocupado mi lugar en el mundo. Al no darse cuenta del cambio usted ofrece los mil dólares a mi réplica y ella acuerda en el trato con la misma satisfacción que yo mostraría. Sin embargo, cuando llega el momento de transferir la propiedad, la conducta de mi réplica y la mía difieren. Ella firma los documentos del caso tal como yo haría y la firma podría convencer a un experto calígrafo. Sin

embargo, mi réplica no le vende a usted el viejo parachoques. No puede hacerlo, dado que no es su dueño. Parecería pues que es falso que mi réplica y yo nos comportaríamos de manera idéntica. Yo vendería el cascajo; ella no podría.

Creo que la manera correcta de responder a esta objeción es conceder el punto. Si estamos dispuestos a privilegiar al ámbito total de las descripciones de sentido común de la conducta, entonces es falso que una persona y su réplica se comportarán siempre de manera idéntica. Sin embargo, no deberíamos suponer que una teoría psicológica va a predecir o a explicar la conducta bajo toda y cualquier descripción que se apoye en el sentido común. Una analogía con la química permitirá ver esto con más claridad. Puede ser verdad que hacer hervir una botella de Chateau Lafitte cause una reducción substancial en su valor de mercado. Pero esto no es algo que esperamos que la química explique por sí sola. Lo que esperamos, en cambio, es que la química explique los efectos de la ebullición, describiéndolos en un vocabulario químico correcto adecuadamente delimitado. Más aún, probablemente no existirá ninguna especificación obvia anterior del ámbito de las descripciones apropiadas de los *explananda* de la química. La elaboración o delimitación del lenguaje en el que los *explananda* van a ser descriptos, es un aspecto, a menudo un aspecto fundamental, de la construcción de teorías científicas (Shapere, 1982). Para explicar por qué la ebullición causa una disminución en el valor de mercado del Chateau Lafitte tendremos que complementar la explicación química de los efectos de la ebullición con hechos acerca de la manera en que los cambios químicos afectan las cualidades sensoriales de un vino y acerca del valor de mercado de los vinos excepcionales de Burdeos. La situación es similar en la psicología. No podemos pretender que la psicología científica explique los eventos conductuales bajo todas las descripciones imaginables. Más bien, el psicólogo tiene que seleccionar o formular un lenguaje descriptivo apropiado para sus *explananda*. Y la formulación de tal vocabulario será una parte fundamental en la construcción de la teoría psicológica.[1]

1. Fodor advierte el punto de manera correcta: "Merece enfatizarse que el sentido de 'conducta' involucra corrección [*is proprietary*], y que eso es lo que tenemos que esperar que ocurra. No toda descripción verdadera de un acto puede ser tal que una teoría de la causación de la conducta explicará al acto bajo esa descripción... Uno no puede tener explicaciones de todo bajo cualquier descripción, y es materia de determinación empírica cuáles son las descripciones de la conducta que revelan su sistematicidad *vis-à-vis* sus causas" (Fodor, 1980a, págs. 330-331; cf. Shapere, 1982).

¿Adónde nos conduce el principio de autonomía? Bien, el argumento del reemplazo sostiene que un organismo y su réplica se comportarán de idéntica manera y que deberán ser considerados psicológicamente idénticos. Pero hemos concedido que un organismo y su réplica no se comportarán de manera idéntica bajo algunas caracterizaciones de su conducta. Permítaseme introducir el término *descripción conductual autonómica* [*autonomous behavioral description*] para [referir a] cualquier descripción de la conducta que satisfaga la siguiente condición: si se aplica a un organismo en una situación [*setting*] dada, entonces se aplicará en esa situación a cualquier réplica del organismo. Parecería, entonces, que la cuestión con la que nos vemos confrontados es la de si las descripciones conductuales autonómicas incluyen todo aquello que un psicólogo encontrará útil para la construcción de las explicaciones sistemáticas de la conducta. Si la respuesta es afirmativa, entonces el argumento del reemplazo nos conduce a la conclusión deseada, dado que las réplicas se comportarán de idéntica manera en situaciones idénticas, cuando la conducta es descripta en el lenguaje descriptivo-conductual [*behavioral-descriptive language*] correcto de los psicólogos. De tal manera, preguntémonos si hay alguna razón para pensar que las descripciones conductuales autonómicas incluyen todo aquello que un psicólogo puede encontrar útil.

Al considerar esta cuestión resulta provechoso reflexionar acerca de la analogía entre los organismos y los robots industriales. Tanto unos como otros son sistemas complejos que en gran medida están controlados internamente y que interactúan con sus respectivos ambientes de maneras sistemáticas. A menos que uno se vea tentado por el dualismo, tiene plausibilidad pensar que las teorías que explican la conducta de distintas clases de robots y las teorías que explican la conducta de distintas clases de organismos, serán aproximadamente análogas. Preguntémonos, pues, si deberíamos esperar que una teoría de la "psicología de los robots" se proponga ofrecer explicaciones de la conducta de los robots bajo descripciones no autonómicas [*non autonomous*

Wilkes formula un punto similar: "Toda ciencia tiene que diseñar una taxonomía de los eventos que caen dentro del dominio de su discurso y, así, tiene que diseñar un vocabulario descriptivo de predicados observacionales y teóricos. Dado que los eventos pueden ser descriptos de variadas maneras, no toda descripción de una acción o de una aptitud para la acción será una descripción en el dominio de la psicología" (Wilkes, 1981, pág. 150).

descriptions]. Como primera observación debemos advertir que hay muchas maneras en que las actividades de un robot podrían ser descriptas en un lenguaje no autonómico. Por ejemplo, un robot de la línea de producción de la General Motors podría, en cierta ocasión, llevar a cabo con éxito su millonésima soldadura autógena. Aunque 'llevar a cabo su millonésima soldadura autógena' podría ser una descripción correcta de lo que hace el robot, no es —claramente— una descripción autonómica. Si antes de realizar la soldadura el robot en cuestión fuera reemplazado por una réplica flamante, la réplica realizaría una soldadura cualitativamente idéntica. Pero estaría llevando a cabo con éxito su primera, no su millonésima soldadura. Al realizar la soldadura un robot podría estar refutando, además, la predicción del profesor Hubert de que ningún robot realizaría jamás un millón de soldaduras y, de manera simultánea, podría estar cumpliendo con una cláusula del contrato entre la General Motors y el fabricante del robot. Pero, nuevamente, ninguna de estas descripciones del robot es autonómica. Parece obvio que si buscamos generalizaciones sistemáticas para explicar la conducta del robot, no deberíamos tener la expectativa de que nuestras generalizaciones expliquen la conducta del robot bajo *esas* descripciones. Las descripciones bajo las cuales esperamos que una teoría de la conducta de los robots explique tal conducta, son descripciones autonómicas.

Nada de esto pretende sugerir que haya algo misterioso respecto del hecho de que el robot más antiguo ha realizado su soldadura millonésima o de que ha refutado la predicción del profesor Hubert o de que ha cumplido con el contrato. Lo que sugiere es que esos hechos y las descripciones que dan cuenta de ellos, se ven mejor como híbridos lógicos o conceptuales. Para llevar a cabo con éxito su soldadura millonésima, el robot tiene que haber realizado una soldadura *y* tiene que haber realizado con anterioridad 999.999 soldaduras. El primer elemento de esa conjunción describe la conducta autonómicamente. Es, precisamente, el tipo de hecho que esperamos que sea explicado por una teoría de la conducta del robot. El segundo elemento de la conjunción es un hecho histórico, y no es el tipo de hecho que esperamos que sea explicado por una teoría de la conducta del robot. Podemos dar una descripción análoga del hecho de que el robot cumplió con una cláusula del contrato. Aquí tenemos, nuevamente, un híbrido conceptual, con un elemento que es el acaecimiento de un evento conductual descripto autonómicamente y otro elemento que es la existencia de un contrato que contiene ciertas cláusulas acerca de lo que el robot hará o de lo que tenga que hacer. Si buscamos un conjunto de generalizaciones que expli-

quen la conducta del robot, no sería razonable esperar que explique este último hecho o el híbrido del que forma parte.

[...] Ahora bien, si la analogía de los robots con los organismos es buena —y creo que lo es—, lo que sugiere es que debemos buscar un patrón de explicación paralelo en la psicología real (como distinta de la "psicología de los robots"). Deberíamos pretender que nuestra teoría se proponga explicar eventos conductuales descriptos autonómicamente. Las descripciones no autonómicas de los eventos conductuales deberían verse como conceptualmente complejas, resolubles en un componente autonómico y un popurrí de factores que explican por qué el evento descripto autonómicamente *cuenta* en tanto satisface la descripción no autonómica. Por supuesto que los demás factores que entran en el análisis de las descripciones conductuales no autonómicas serán bastante más ricos y más complejos cuando los sujetos de nuestra teoría sean personas en vez de ser animales o robots. Esos factores pueden incluir la historia del individuo en cuestión, la historia de los términos que usa, las prácticas lingüísticas, sociales, legales y rituales que se dan en la sociedad de la que forma parte, y también quizá muchos otros más. Así, si nuestra analogía es buena, es plausible concluir que las descripciones de la conducta que una teoría psicológica debería usar en sus *explananda* serían descripciones autonómicas. Ésta es, justamente, la conclusión que precisamos para dar ímpetu al argumento del reemplazo y, así, dar apoyo al principio de autonomía.

Para el punto de vista que estoy defendiendo, las descripciones de sentido común no autonómicas de la conducta son típicamente híbridos conceptuales. A veces dispondremos de una descripción de sentido común del componente autonómico de un acto no autonómico. Pero no es necesario que siempre sea así. Puede resultar que sea necesario efectuar un trabajo substancial para forjar descripciones conductuales autonómicas apropiadas para usar en la psicología científica (véase Alston, 1974). Pero esto no debe sorprender porque, como señalé antes, la formulación de una terminología apropiada para describir los *explananda* es a menudo un paso esencial en el desarrollo de una ciencia nueva. Dicho sea de paso, adviértase que no hay razón para suponer que la terminología autonómica descriptiva de la conducta, que en última instancia se considera la más útil, sea una descripción puramente física de los movimientos del tipo que los conductistas buscaron y que nunca hallaron.

Al desarrollar descripciones conductuales no autonómicas híbridas, el sentido común produce descripciones conductuales que son de grano

mucho más fino que el de las que dispondríamos si nos restringiéramos a descripciones autonómicas. No hay nada irrazonable en esto, dado que a menudo nuestras preocupaciones prácticas exigen una descripción más fina de la conducta. Pero, si estoy en lo correcto, entonces esas preocupaciones prácticas nos conducen a una taxonomía de la conducta que no cuadra con una ciencia sistemática que se proponga explicar a la conducta. La psicología de sentido común ha seguido la estrategia del sentido común al desarrollar un conjunto de descripciones híbridas para los *estados mentales*, que incorporan variados rasgos históricos, contextuales y comparativos del organismo. De tal modo... la noción de sentido común de creer que p es una amalgama de consideraciones históricas, contextuales, ideológicas y quizá de otro tipo. Sin duda que esta manera de "desmenuzar" el bollo mental ha probado ser eficiente y útil para la tarea diaria de tratar con otras personas. Si no fuera así, no habría sobrevivido. Sin embargo, el núcleo crítico del principio de autonomía es que al incorporar en las descripciones de los estados mentales rasgos históricos, contextuales e ideológicos, la psicología de sentido común ha hecho una taxonomía de los estados demasiado estrecha, trazando distinciones que resultan ser innecesarias e inmanejables cuando se trata de buscar una explicación de la conducta sistemática y causal. Creer que p es estar en un estado autonómico funcional *y* tener con el adscriptor [*ascriber*] una cierta historia, un contexto y una relación ideológica. Estos factores adicionales pueden ser estudiados, seguramente, por distintas disciplinas. Pero ellos no tienen cabida en una ciencia que se proponga explicar la conducta. Al "desmenuzar" el bollo de manera demasiado fina, tales factores impiden la formulación de las generalizaciones que se aplican por igual a un organismo y a su réplica.

La TRM Fuerte nos hará expresar las generalizaciones cognitivas en el lenguaje híbrido de la adscripción de contenido. La Teoría Sintáctica de la Mente, en cambio, requiere generalizaciones puramente formales que ignoran los factores históricos y ambientales que pueden distinguir a un organismo de su réplica a los ojos de la psicología de sentido común. Si el argumento en favor del principio de autonomía es persuasivo, entonces la estrategia de la TSM es la que corresponde preferir.

TRADUCTORES: Eleonora Baringoltz y Eduardo Rabossi.

REFERENCIAS BIBLIOGRÁFICAS

Alston, W.P.: (1974) "Conceptual prolegomena to a psychological theory of intentional action", en S.C. Brown (comp.), *Philosophy of Psychology*, Harper & Row.

Fodor, J.A.: (1974) "Special sciences", *Synthèse* 28, 97-115.

Fodor, J.A.: (1978) "Propositional attitudes", *The Monist* 61, 4, reimpreso en Fodor (1981a).

Fodor, J.A.: (1980a) "Methodological solipsism considered as a research strategy in cognitive psychology", *Behavioral and Brain Sciences* 3, 63-73, reimpreso en Fodor (1981a).

Fodor, J.A.: (1980b) "Methodological solipsism: Replies to commentators", *Behavioral and Brain Sciences* 3, 99-108.

Fodor, J.A.: (1981a) *Representations*, Cambridge, Mass., MIT Press.

Fodor, J.A.: (198lb) "Introduction: Something on the state of the art", reimpreso en Fodor (1981a).

Kim, J.: (1969) "Events and their descriptions: Some considerations", en N. Rescher *et al.* (comps.), *Essays in Honor of C. G. Hempel*, Dordrecht, Reidel.

Kim, J.: (1976) "Events and property exemplifications", en M. Brand y D. Walton (comps.), *Action Theory*, Dordrecht, Reidel.

Kim, J.: (1978) "Supervenience and nomological incommensurables", *American Philosophical Quarterly* 15, 149-156.

Kim, J.: (1982) "Psychological supervenience", *Philosophical Studies* 41, 51-70.

Putnam, H.: (1975) "The meaning of 'meaning' ", en K. Gunderson (comp.) *Minnesota Studies in the Philosophy of Science, vol. 7: Language. Mind and Knowledge*, University of Minnesota Press.

Shapere, D.: (l982) "The concept of observation in science and philosophy", *Philosophy of Science*, 49, 485-525.

Stich, S.: (1978) "Autonomous psychology and the belief-desire thesis", *The Monist*, 6l, 573-91.

Wilkes, K.: (1981) "Functionalism, psychology and the philosophy of mind", *Philosophical Topics*, l2, 147-l67.

V

EL SIGNIFICADO Y LOS CONTENIDOS MENTALES

A. Los *Doppelgänger* y el anti-individualismo

CAPÍTULO 8

SIGNIFICADO Y REFERENCIA*

Hilary Putnam

Poco clara como es, la doctrina tradicional de que la noción de "significado" posee la ambigüedad extensión/intensión, tiene ciertas consecuencias típicas. La doctrina de que el significado de un término es un concepto, acarrea la implicación de que los significados son entidades mentales. Frege, no obstante, se rebeló contra este "psicologismo". Pensando que los significados son propiedad pública —que el *mismo* significado puede ser "captado" ["*grasped*"] por más de una persona y por las mismas personas en momentos diferentes—, identificó a los conceptos (y aun a las "intensiones" o a los significados) con entidades abstractas antes que con entidades mentales. No obstante, el "captar" esas entidades abstractas era todavía un acto psicológico individual. Ninguno de estos filósofos dudó de que comprender una palabra (conocer su intensión) era cuestión de estar en un cierto estado psicológico (de algún modo a la manera en que saber cómo factorear números mentalmente [*in one's head*] es justamente estar en un cierto estado psicológico muy complejo).

En segundo lugar, el conocido ejemplo de los dos términos 'criatura con riñón' y 'criatura con corazón' muestra que dos términos pueden tener la misma extensión y no obstante diferir en intensión. Pero se consideró obvio que la inversa es imposible: dos términos no pueden diferir en extensión y tener la misma intensión. Es interesante que ningún argumento en favor de esta imposibilidad fue alguna vez ofrecido. Probablemente refleja la tradición de los filósofos antiguos y medievales que supusieron que el concepto correspondiente a un término era una conjunción de predicados y, en consecuencia, que el concepto correspondiente a un término tiene *siempre* que proveer una condición necesaria y suficiente para caer en la extensión del término. Para filósofos como Carnap, que aceptó la teoría verificacionista del significado, el

* *The Journal of Philosophy* 70 (1973), 699-711. Con autorización del autor y del *Journal of Philosophy*.

concepto correspondiente a un término proveía (en el caso ideal, donde el término tenía "significado completo") un *criterio* de pertenencia a la extensión (no en el sentido de "condición necesaria y suficiente", sino en el sentido fuerte de *modo de reconocer* si una cosa dada cae dentro de la extensión o no). De este modo la teoría del significado vino a descansar en dos presupuestos indiscutidos:

(1) Que conocer el significado de un término es estar en un cierto estado psicológico (en el sentido de "estado psicológico", según el cual los estados de recuerdo [*memory*] y de creencia son "estados psicológicos"; por supuesto nadie pensó que conocer el significado de una palabra fuera un estado de conciencia continuo).

(2) Que el significado de un término determina su extensión (en el sentido de que mismidad [*sameness*] de intensión implica mismidad de extensión).

Argumentaré que estas dos presuposiciones no son satisfechas conjuntamente por *ninguna* noción, no ya por ninguna noción de significado. El concepto tradicional de significado es un concepto que descansa en una teoría falsa.

¿Están los significados en la cabeza?

A los efectos de los siguientes ejemplos de ciencia ficción supondremos que hay en algún lugar un planeta que llamaremos Tierra Gemela. La Tierra Gemela es muy parecida a la Tierra: de hecho la gente en la Tierra Gemela incluso habla en *español*. De hecho, aparte de las diferencias que especificaremos en nuestros ejemplos de ciencia ficción, el lector puede suponer que la Tierra Gemela es *exactamente* como la Tierra. Puede aun suponer si lo desea, que tiene un *Doppelgänger* —una copia idéntica— en la Tierra Gemela, a pesar de que mis relatos no dependerán de eso.

A pesar de que alguna gente en la Tierra Gemela (digamos, aquellos que se llaman a sí mismos "argentinos" y aquellos que se llaman a sí mismos "uruguayos" y aquellos que se llaman a sí mismos "españoles", etcétera) habla español, no es sorprendente, que haya unas diferencias mínimas entre los dialectos de los hablantes del español en la Tierra Gemela y el español estándar.

Una de las peculiaridades de la Tierra Gemela es que el líquido llamado "agua" no es H_2O sino un líquido diferente cuya fórmula química es muy larga y complicada. Abreviaré esta fórmula química simplemen-

te como XYZ. Supondré que XYZ es indistinguible del agua a temperaturas y presiones normales. Supondré también que los océanos, lagos y mares de la Tierra Gemela contienen XYZ y no agua, que en la Tierra Gemela llueve XYZ y no agua, etcétera.

Si una nave espacial de la Tierra visita alguna vez la Tierra Gemela, la primera suposición será que "agua" tiene el mismo significado en la Tierra y en la Tierra Gemela. Esta suposición será corregida cuando se descubra que "agua" en la Tierra Gemela es XYZ, y la nave espacial terráquea reportará algo como lo que sigue:

"En la Tierra Gemela la palabra "agua" significa XYZ".

Simétricamente, si una nave espacial de la Tierra Gemela visita alguna vez la Tierra, la primera suposición será que la palabra "agua" tiene el mismo significado en la Tierra Gemela y en la Tierra. Esta suposición será corregida cuando se descubra que "agua" en la Tierra es H_2O, y la nave espacial de la Tierra Gemela reportará:

"En la Tierra la palabra "agua" significa H_2O".

Nótese que no hay ningún problema con la extensión del término 'agua': simplemente la palabra tiene dos significados diferentes (como se dice); en el sentido en que es usada en la Tierra Gemela, el sentido de $agua_{TG}$, lo que *nosotros* llamamos "agua" simplemente no es agua, mientras que en el sentido en que es usada en la Tierra, el sentido de $agua_T$, lo que los habitantes de la Tierra Gemela llaman "agua" simplemente no es agua. La extensión de "agua" en el sentido de $agua_T$ es el conjunto de las totalidades que consisten de las moléculas H_2O, o algo parecido; la extensión de 'agua' en el sentido de $agua_{TG}$ es el conjunto de las totalidades que consisten de las moléculas XYZ, o algo parecido.

Dejemos ahora que el tiempo retroceda hacia alrededor de 1750. El hablante terrestre típico del español no sabía que el agua consistía de hidrógeno y oxígeno, y el hablante típico del español de la Tierra Gemela no sabía que "agua" consistía de XYZ. Sea $Oscar_1$ ese típico hablante terráqueo del español, y sea $Oscar_2$ su contraparte en la Tierra Gemela. Se puede suponer que no hay ninguna creencia que $Oscar_1$ tenga acerca del agua que $Oscar_2$ no tenga acerca de "agua". Si se quiere puede suponerse aun que $Oscar_1$ y $Oscar_2$ fueron duplicados exactos en apariencia, sentimientos, pensamientos, monólogo interior, etcétera. Sin embargo, la extensión del término 'agua' era H_2O en la Tierra tanto en 1750 como

en 1950; y la extensión del término 'agua' era XYZ en la Tierra Gemela en 1750 como en 1950. Oscar$_1$ y Oscar$_2$ entendían el término 'agua' en forma diferente en 1750 *a pesar de que estaban en el mismo estado psicológico*, y a pesar de que, dado el estado de la ciencia de la época, le haya tomado a sus comunidades científicas alrededor de cincuenta años descubrir que ellos entendían el término 'agua' de manera diferente. Así, la extensión del término 'agua' (y, de hecho, su "significado" en el uso preanalítico intuitivo de este término) *no* es una función del estado psicológico del hablante por sí mismo.[1]

Pero, podría objetarse, ¿por qué deberíamos aceptar que el término 'agua' tuviera la misma extensión en 1750 y en 1950 (en ambas Tierras)? Supóngase que señalo un vaso con agua y digo "Este líquido se llama agua". Mi "definición ostensiva" de agua tiene la siguiente presuposición empírica: que la porción de líquido a la que señalo tiene una cierta relación de mismidad (digamos, *x es el mismo líquido que y*, o *x es el* $mismo_L$ *que y*) con la substancia [*stuff*] a la que yo y otros hablantes en mi comunidad lingüística, en otras ocasiones, hemos llamado "agua". Si esta presuposición es falsa porque, digamos, estoy señalando sin saberlo, a un vaso con ginebra y no a uno con agua, entonces no pretendo que mi definición ostensiva sea aceptada. Así, la definición ostensiva conlleva lo que podría ser llamada una condición necesaria y suficiente "revocable" ["*defeasible*"]: la condición necesaria y suficiente para ser agua es tener la relación $mismo_L$ con lo que está en el vaso, pero ella es la condición necesaria y suficiente sólo si la presuposición empírica es satisfecha. Si no lo es, entonces una de unas series de, por decir así, condiciones "de resguardo" ["*fallback*"] se activan.

El punto clave es que la relación $mismo_L$ es una relación *teórica*: determinar que algo sea o no sea el mismo líquido que *éste* puede requerir un monto indeterminado de investigación científica. Así, el hecho de que en 1750 un hablante del español pudiera haber llamado a XYZ "agua", mientras que él o sus sucesores no habrían llamado XYZ al agua en 1800 o en 1850, no significa que el "significado" de 'agua' cambiara en el intervalo para el hablante promedio. En 1750 o en 1850 o en 1950 uno podría haber señalado, digamos, al líquido en el lago Michigan como un ejemplo de "agua". Lo que cambió fue que en 1750 habríamos pensado equivocadamente que XYZ tenía la relación $mismo_L$ con el líquido en el lago Michigan, mientras que en 1800 o en 1850 habríamos sabido que no era así.

1. Ver nota 2 y el texto correspondiente.

Permítasenos ahora modificar nuestro relato de ciencia ficción. Supondré que los cacharros y las cacerolas de molibdeno *no pueden* ser distinguidos de los cacharros y las cacerolas de aluminio, salvo por un experto. (Por lo que sé, esto podría ser verdadero, y a fortiori, podría ser verdadero, por lo que sé, en virtud de "conocer el significado" de las palabras *aluminio* y *molibdeno*). Ahora supondremos que el molibdeno es tan común en la Tierra Gemela como lo es el aluminio en la Tierra, y que el aluminio es tan raro en la Tierra Gemela como el molibdeno lo es en la Tierra. En particular, supondremos que en la Tierra Gemela los cacharros y cacerolas de "aluminio" están hechos de molibdeno. Finalmente, supondremos que las palabras 'aluminio' y 'molibdeno' están *intercambiadas* en la Tierra Gemela: 'aluminio' es el nombre del *molibdeno*, y 'molibdeno' es el nombre del *aluminio*. Si una nave espacial de la Tierra visitara la Tierra Gemela, los visitantes terráqueos no sospecharían que en la Tierra Gemela los cacharros y cacerolas de "aluminio" no estaban hechos de aluminio, especialmente cuando los habitantes de la Tierra Gemela *dijeran* que lo estaban. Pero hay una diferencia importante entre los dos casos. Un metalúrgico terráqueo podría reconocer fácilmente que el "aluminio" era molibdeno, y un metalúrgico de la Tierra Gemela podría, del mismo modo, reconocer fácilmente que el aluminio era "molibdeno". (Las desagradables comillas en el párrafo precedente indican los usos de los habitantes de la Tierra Gemela.) Mientras que en 1750 nadie en la Tierra ni en la Tierra Gemela podría haber distinguido el agua del "agua", la confusión de aluminio con "aluminio" involucra a sólo una parte de las comunidades lingüísticas en cuestión.

Este ejemplo apunta a lo mismo que el precedente. Si Oscar$_1$ y Oscar$_2$ son respectivamente, hablantes estándar del español terrestre y del español de la Tierra Gemela, y ninguno tiene un conocimiento químico o metalúrgico sofisticado, entonces no puede haber ninguna diferencia en sus estados psicológicos cuando usan la palabra 'aluminio'; no obstante, debemos decir que 'aluminio' tiene la extensión *aluminio* en el idiolecto de Oscar$_1$ y la extensión *molibdeno* en el idiolecto de Oscar$_2$. (También tenemos que decir que Oscar$_1$ y Oscar$_2$ quieren decir [*mean*] cosas diferentes con 'aluminio'; que 'aluminio' tiene un significado diferente en la Tierra del que tiene en la Tierra Gemela, etcétera.) Nuevamente, vemos que el estado psicológico del hablante *no* determina la extensión (*o* el "significado", hablando preanalíticamente) de la palabra.

Antes de seguir discutiendo este ejemplo introduzcamos un ejemplo

que *no* es de ciencia ficción. Supongamos que usted como yo no puede distinguir un olmo de una haya. Sin embargo, decimos que la extensión de 'olmo' en mi idiolecto es la misma que la extensión de 'olmo' en el de cualquiera otra persona, a saber, el conjunto de todos los olmos, y que el conjunto de todas las hayas es la extensión de 'haya' en *ambos* idiolectos. Así, 'olmo' en mi idiolecto tiene una extensión diferente de la de 'haya' en su idiolecto (tal como debe ser). ¿Es realmente creíble que esta diferencia en la extensión la provoque alguna diferencia en nuestros *conceptos*? Mi *concepto* de un olmo es exactamente el mismo que mi concepto de un haya (me sonrojo al confesarlo). Si alguien intentase heroicamente sostener que la diferencia entre la extensión de 'olmo' y la extensión de 'haya' en *mi* idiolecto se explica por una diferencia en mi estado psicológico, entonces podríamos refutarlo construyendo un ejemplo del tipo Tierra Gemela, intercambiando simplemente las palabras 'olmo' y 'haya' en la Tierra Gemela (como se hizo en el ejemplo anterior de 'aluminio' y de 'molibdeno'). Más aún, supongamos que en la Tierra Gemela tengo un *Doppelgänger* molécula por molécula "idéntico" a mí. Si se es dualista, supóngase también que mi *Doppelgänger* tiene los mismos pensamientos verbalizados que yo, los mismos datos sensoriales, las mismas disposiciones, etcétera. Es absurdo pensar que *su* estado psicológico sea diferente del mío: sin embargo, él "quiere decir" *haya* cuando dice 'olmo' y *yo* "quiero decir" *olmo* cuando digo olmo. Míreselo como se lo mire, ¡los "significados" no están en la *cabeza*!

Una hipótesis sociolingüística

Los últimos dos ejemplos dependen de un hecho acerca del lenguaje que, sorprendentemente, parece no haber sido señalado nunca: que hay una *división del trabajo lingüístico*. Difícilmente podríamos usar palabras tales como 'olmo' y 'aluminio' si no tuviésemos una manera de reconocer los árboles olmos y el metal aluminio, pero no todos aquellos para quienes la distinción es importante, tienen que saber hacerla. Cambiemos de ejemplo: consideremos el *oro*. El oro es importante por muchas razones: es un metal precioso, es un metal monetario, tiene un valor simbólico (es importante para la mayoría de la gente que su anillo matrimonial de "oro" sea *realmente* hecho de oro y no sólo *parezca* de oro), etcétera. Consideremos a nuestra comunidad como una "fábrica": en esta "fábrica" algunas personas tienen la "tarea" de usar *anillos*

matrimoniales de oro, otras tienen la "tarea" de vender anillos matrimoniales de oro y otras tienen la "tarea" de *determinar si algo realmente es oro o no lo es*. No es de manera alguna necesario o ni siquiera útil que quien use un anillo de oro (o gemelos de oro, etcétera) o discuta el "patrón oro", etcétera, tenga que ver con la compra y venta de oro. Ni tampoco es necesario o útil que quienquiera que compre o venda oro sea capaz de decir si algo es realmente oro o no lo es, en una sociedad donde esa deshonestidad (vender oro falso) no sea común y en la cual uno pueda fácilmente consultar a un experto en caso de duda. Y no es *ciertamente* necesario ni útil que quienquiera que tenga la ocasión de comprar o usar oro sepa determinar con algún grado de seguridad si algo es o no realmente oro.

Los hechos precedentes son ejemplos (en un sentido amplio) de la división mundana del trabajo. Pero engendran una división de las tareas lingüísticas: toda persona para la cual el oro es importante por alguna razón tiene que *adquirir* [*acquire*] la palabra 'oro', pero no tiene que aprender el *método de reconocer* si algo es oro o no lo es. Puede fiarse de una subclase especial de hablantes. Las características que generalmente se piensan presentes en conexión con un nombre general —condiciones necesarias y suficientes para la pertenencia a la extensión, modos de reconocer si algo está en la extensión, etcétera— están todas presentes en la comunidad lingüística *considerada como un cuerpo colectivo*, pero este cuerpo colectivo divide la "tarea" de conocer y la de emplear las distintas partes del "significado" de 'oro'.

Por supuesto que esta división del trabajo lingüístico descansa sobre la división del trabajo *no* lingüístico, y la presupone. Si sólo los que saben cómo distinguir si un cierto metal es oro o no lo es, tuviesen alguna razón para tener la palabra 'oro' en su vocabulario, entonces la palabra 'oro' sería como era la palabra 'agua' en 1750 con respecto a esa subclase de hablantes, y los otros hablantes no la habrían adquirido. Y algunas palabras no exhiben ninguna división de trabajo lingüístico: 'silla', por ejemplo. Pero con el incremento de la división del trabajo en la sociedad y el surgimiento de la ciencia, más y más palabras comienzan a exhibir este tipo de división del trabajo. 'Agua', por ejemplo, no exhibía esa división, antes del surgimiento de la química. En nuestros días es necesario, obviamente, para cualquier hablante ser capaz de reconocer el agua (de manera confiable bajo condiciones normales), y probablemente la mayoría de los hablantes adultos saben incluso que el "agua es H_2O" es la condición necesaria y suficiente, pero sólo unos pocos hablantes adultos podrían distinguir el agua de los líquidos que se le

parecen superficialmente. En caso de duda, otros hablantes se fiarían del juicio de esos hablantes "expertos". Así, el modo de reconocer poseído por esos hablantes "expertos" es también poseído, a través de ellos, por el cuerpo lingüístico colectivo, aun cuando no lo posea cada miembro individual, y de esta manera el hecho más *investigado* acerca del agua puede llegar a ser parte del significado *social* de la palabra, aunque desconocido por casi todos los hablantes que adquieren la palabra.

Me parece que para los sociolingüistas será muy importante investigar este fenómeno de la división del trabajo lingüístico. En conexión con él, quiero proponer la siguiente hipótesis:

> HIPÓTESIS DE LA UNIVERSALIDAD DE LA DIVISIÓN DEL TRABAJO LINGÜÍSTICO: Toda comunidad lingüística ejemplifica la clase de división del trabajo lingüístico descripto, esto es, posee al menos algunos términos cuyos "criterios" asociados son conocidos sólo por un subconjunto de los hablantes que adquieren los términos; y cuyo uso por parte de otros hablantes depende de una cooperación estructurada entre ellos y los hablantes de los subconjuntos relevantes.

Es fácil ver cómo este fenómeno da cuenta de algunos de los ejemplos dados más arriba relativos al fracaso de los supuestos (1 y 2). Cuando un término está sujeto a la división del trabajo lingüístico, el hablante "medio" que lo adquiere no adquiere nada que fije su extensión. En particular, su estado psicológico individual no fija, *ciertamente*, su extensión; es sólo el estado sociolingüístico del cuerpo lingüístico colectivo al cual pertenece el hablante, el que fija la extensión.

Podríamos resumir esta discusión señalando que en el mundo hay dos clases de herramientas: hay herramientas como un martillo o un destornillador que pueden ser usadas por una persona; y hay herramientas como un barco de vapor que requieren para su uso la actividad cooperativa de un cierto número de personas. Las palabras han sido concebidas con demasiado apego al modelo del primer tipo de herramientas.

Deicticidad [indexicality] *y rigidez*

El primero de nuestros ejemplos de ciencia ficción —'agua' en la Tierra y en la Tierra Gemela en 1750— no involucra una división del trabajo lingüístico, o por lo menos no lo involucra de la misma manera que los ejemplos de 'aluminio' y de 'olmo'. En la Tierra no había ningún

'experto' en agua en 1750 (al menos en nuestro relato), ni había expertos en "agua" en la Tierra Gemela. El ejemplo involucra cosas, que ahora discutiremos, que son de importancia fundamental para la teoría de la referencia y también para la teoría de la verdad necesaria.

Sean M_1 y M_2 dos mundos posibles en los cuales yo existo y en los cuales este vaso existe y en los cuales doy una explicación del significado señalando a este vaso y diciendo: "Esto es agua". Supongamos que en M_1 el vaso está lleno de H_2O y que en M_2 el vaso está lleno de XYZ. También supongamos que M_1 es el mundo *real* y que XYZ es el material típicamente llamado 'agua' en el mundo M_2 (de tal manera que la relación entre los hablantes del español en M_1 y los hablantes del español en M_2 es exactamente la misma que la relación entre los hablantes del español en la Tierra y los hablantes del español en la Tierra Gemela). Entonces, uno podría tener dos teorías acerca del significado de 'agua':

(1) Uno podría sostener que 'agua' es *relativa-a-los-mundos* [*world-relative*] aunque *constante* en significado (esto es, la palabra tiene un significado relativo *constante*). En esta teoría, 'agua' *significa lo mismo* en M_1 y M_2; es decir, agua es precisamente H_2O en M_1 y agua es XYZ en M_2.

(2) Uno podría sostener que agua es H_2O en todos los mundos (el material llamado "agua" en M_2 no es agua), pero 'agua' no tiene el mismo significado en M_1 y M_2.

Si lo que se ha dicho antes acerca de la Tierra Gemela es correcto, entonces (2) es claramente correcta. Cuando digo "*Este* (líquido) es agua", el "este" es, por así decirlo, un "este" *de re*, es decir, la fuerza de mi explicación proviene de que "agua" es toda cosa que cumpla una cierta relación de equivalencia (la relación que antes llamamos "$mismo_L$" con la porción de líquido a la que uno se refiere como "este" *en el mundo real*.

Podríamos simbolizar la diferencia entre las dos teorías como una diferencia de "alcance" ["*scope*"], de la siguiente manera. En la teoría 1, lo siguiente es verdadero:

> (1') (Para todo mundo M) (Para todo x en M) (x es agua $\equiv x$ tiene [la relación] $mismo_L$ con la entidad referida como "esto" en M)

mientras que en la teoría (2):

(2') (Para todo mundo *M*) (Para todo *x* en *M*) (*x* es agua ≡ *x* tiene [la relación] $mismo_L$ con la entidad referida como "esto" *en el mundo real* M_1)

Llamo a eso una diferencia de "alcance" porque en (1') 'la entidad referida como "esto" está dentro del alcance de 'Para todo mundo *M*', como hace explícito la frase calificativa 'en *M*'; mientras que en (2') 'la entidad referida como "esto"' significa [*means*] "la entidad referida como 'esto' en el mundo real", y tiene así una referencia *independiente* de la variable ligada '*M*'.

Kripke llama "rígido" (en una oración dada) a un designador [*designator*] si (en esa oración) refiere al mismo individuo en todo mundo posible en el cual el designador designa. Si extendemos esta noción de rigidez a los nombres de substancias, entonces podríamos expresar la teoría de Kripke y la mía diciendo que el término "agua" es *rígido*.

La rigidez del término "agua" se sigue del hecho de que, cuando doy la definición ostensiva "*Este* (líquido) es agua", intento significar (2') y no (1').

También podríamos decir, siguiendo a Kripke, que cuando doy la definición ostensiva "*Este* (líquido) es agua", el demostrativo 'este' es *rígido*.

Kripke fue el primero en observar que esta teoría del significado (o del "uso", o de lo que fuera) de la palabra 'agua' (y también de otros términos de clase natural) tiene consecuencias sorprendentes para la teoría de la verdad necesaria.

A fin de explicar esto, permítaseme introducir la noción de "relación de entre-mundos" [*cross-world relation*]. Una relación binaria R se llamará "entre-mundos" cuando se la comprenda de un modo tal que su extensión sea un conjunto de pares ordenados de individuos *que no están todos en el mismo mundo posible*. Por ejemplo, es fácil comprender la relación *misma altura que* como una relación entre-mundos: hay que entenderla de modo que, por ejemplo, si *x* es un individuo en un mundo M_1, que mide 1.70 (en M_1), e y es un individuo en M_2 que mide 1.70 (en M_2), entonces el par ordenado *x,y* pertenece a la extensión de *misma altura que*. (Puesto que un individuo puede tener diferente altura en mundos posibles diferentes en los que el mismo individuo exista, estrictamente hablando no es el par ordenado *x,y* el que constituye un elemento de la extensión de *misma altura que*, sino más bien el par ordenado x-*en el mundo*-M_1, y-*en el mundo*-M_2.)

Del mismo modo, podemos comprender la relación $mismo_L$ (el

mismo líquido que) como una relación entre-mundos comprendiéndola de modo tal que un líquido en el mundo M_1 que tenga las mismas propiedades físicas relevantes (en M_1) que las que posee (en M_2) un líquido en M_2, tenga la relación *mismo*$_L$ con el último líquido.

La teoría que hemos venido presentando puede resumirse entonces diciendo que una entidad *x* en un mundo posible cualquiera, es *agua* si y sólo si tiene la relación *mismo*$_L$ (interpretada como una relación entre-mundos) con el material que *nosotros* denominamos "agua" en el mundo real.

Supóngase, ahora, que yo todavía no he descubierto cuáles son las propiedades físicas relevantes del agua (en el mundo real), esto es, que yo no sé todavía que el agua es H_2O. Podría ser que tenga maneras exitosas de *reconocer* el agua (por supuesto, podría cometer un pequeño número de errores que no sería capaz de detectar hasta una etapa ulterior en nuestro desarrollo científico), pero que no conozca la microestructura del agua. Si estoy de acuerdo en que un líquido con las propiedades superficiales del "agua" pero con una microestructura diferente *no es realmente* agua, entonces mis maneras de reconocer el agua no pueden ser vistas como una especificación analítica de lo que *es ser* agua. Más bien, la definición operacional, como la ostensiva, simplemente es una manera de señalar un patrón, señalando la sustancia *en el mundo real* tal que, que x sea agua en *cualquier* mundo equivale a que *x* mantenga la relación *mismo*$_L$ con los miembros *normales* de la clase de entidades *locales* que satisfacen la definición operacional. "Agua" en la Tierra Gemela no es agua, aun si satisface la definición operacional, porque no tiene la relación *mismo*$_L$ con la sustancia local que satisface la definición operacional, y la sustancia local que satisface la definición operacional pero que tiene una microestructura diferente del resto de la sustancia local que satisface la definición operacional, tampoco es agua, porque no mantiene *mismo*$_L$ con las muestras *normales* del "agua" local.

Supóngase ahora que he descubierto la microestructura del agua: que el agua es H_2O. Seré capaz entonces de decir que lo que en la Tierra Gemela antes *tomé equivocadamente* por agua, no es realmente agua. Del mismo modo, si uno describe, no ya otro planeta en el universo real, sino otro universo posible en el cual haya un material con la fórmula química XYZ que pasa el "test operacional" para el *agua*, tendremos que decir que ese elemento no es agua sino meramente XYZ. Uno no habrá descripto un mundo posible en el cual "El agua es XYZ", sino meramente un mundo posible en el cual existen lagos de XYZ, la gente

bebe XYZ (y no agua), o lo que sea. En efecto, una vez que hemos descubierto la naturaleza del agua, no hay nada que cuente como un mundo posible en el cual el agua no tenga esa naturaleza. Una vez que hemos descubierto que el agua (en el mundo real) es H_2O, *no hay nada que cuente como un mundo posible en el cual el agua no sea* H_2O.

Por otro lado, podemos imaginar, perfectamente, tener experiencias que no convencerían (y que harían racional creer) que el agua *no es* H_2O. En este sentido, es concebible que el agua no sea H_2O. Es concebible, ¡pero no es posible! Lo concebible no es una prueba de lo posible.

Kripke alude a los enunciados que racionalmente son no revisables (suponiendo que existan), como *cognoscitivamente necesarios*. Y alude simplemente como necesarios (o a veces como "metafísicamente necesarios") a los enunciados que son verdaderos en todos los mundos posibles. En esta terminología, el punto que acabamos de elaborar puede volver a enunciarse así: un enunciado puede ser (metafísicamente) necesario y cognoscitivamente contingente. La intuición humana no tiene un acceso privilegiado a la necesidad metafísica.

En este artículo, sin embargo, nuestro interés reside en la teoría del significado y no en la teoría de la verdad necesaria. Desde hace tiempo, palabras como 'ahora', 'esto', 'aquí' han sido reconocidas como *deícticos*, o *caso-reflexivas* [*token-reflexive*], esto es, como teniendo una extensión que varía de contexto a contexto o de caso a caso. Para estas palabras, nadie ha sugerido la teoría tradicional de que "la intensión determina la extensión". Para tomar nuestro ejemplo de la Tierra Gemela: si yo tengo un *Doppelgänger* en la Tierra Gemela, entonces cuando pienso "Me duele la cabeza", él piensa "Me duele la cabeza". Pero la extensión del ejemplar particular de 'yo' en su pensamiento verbalizado es él mismo (o su clase unitaria, para ser preciso), mientras que la extensión del ejemplar de 'yo' en *mi* pensamiento verbalizado soy *yo* (o mi clase unitaria, para ser preciso). Así, la misma palabra, 'yo', tiene dos extensiones diferentes en dos idiolectos diferentes; pero no se sigue que el concepto que yo tengo de mí mismo sea diferente del concepto que mi *Doppelgänger* tiene de sí mismo.

Ahora bien, hemos sostenido que la deicticidad se extiende más allá de las palabras y morfemas (por ejemplo, el tiempo de los verbos) *obviamente* deicticales. Nuestra teoría se puede resumir diciendo que las palabras como 'agua' tienen un componente deíctico no advertido: "agua" es el material que tiene una cierta relación de similaridad [*similarity*] con el agua que *nos rodea*. El agua en otra época o en otro lugar o incluso en otro mundo posible, tiene que tener, a fin de que sea agua,

la relación *mismo*$_L$ con *nuestra* "agua". Así, la teoría de que 1) las palabras tienen "intensiones" que son algo así como conceptos que los hablantes asocian con las palabras; y 2) que la intensión determina la extensión, no puede ser verdadera de las palabras de clase natural como 'agua', por la misma razón que no puede ser verdadera de las palabras obviamente deícticas como 'yo'.

La teoría de que las palabras de clase natural como 'agua' son deícticas, deja pendiente sin embargo el problema de decir si 'agua' en el dialecto castellano de la Tierra Gemela tiene el mismo *significado* que 'agua' en el dialecto de la Tierra y una extensión diferente —que es lo que normalmente decimos respecto de 'yo' en los diferentes idiolectos—, renunciando con ello a la doctrina de que "el significado (la intensión) determina la extensión", o si decir, tal como hemos elegido, que la diferencia en la extensión constituye ipso facto una diferencia en el significado de las palabras de clase natural, renunciando con ello a la doctrina de que los significados son conceptos o, ciertamente, entidades mentales de *algún* tipo.[2]

Debería resultar claro, sin embargo, que la doctrina de Kripke de que las palabras de clase natural son designadores rígidos y nuestra doctrina de que son deícticas, no son sino dos modos de establecer el mismo punto.

Hemos visto entonces que la extensión de un término no se fija mediante un concepto que el hablante individual tiene en su cabeza, y esto es verdad tanto porque la extensión se determina, en general, *socialmente*— existe la división del trabajo lingüístico así como la del trabajo "real"— como porque la extensión se determina, en parte, *deícticamente*. La extensión de nuestros términos depende de la naturaleza real de las cosas particulares que sirven como paradigmas, y generalmente, esta naturaleza real no es conocida totalmente por el hablante. La teoría semántica tradicional deja a un lado las dos contribuciones

2. Nuestras razones para rechazar la primer opción —decir que 'agua' tiene el mismo significado en la Tierra y en la Tierra Gemela, renunciando por eso a la doctrina de que el significado determina la referencia—, se presentan en "The Meaning of 'Meaning'". Ellas se pueden ilustrar así: Supóngase que 'agua' tiene el mismo significado en la Tierra y en la Tierra Gemela. Ahora bien, permítase que la palabra 'agua' resulte fonémicamente diferente en la Tierra Gemela, que resulte, digamos, 'quaxel'. Presumiblemente éste no es per se un cambio en el significado, en ningún enfoque. Así, 'agua' y 'quaxel' tienen el mismo significado (aunque se refieran a líquidos diferentes). Pero esto resulta altamente contraintuitivo. ¿Por qué no decir entonces, que 'olmo' en mi idiolecto tiene el mismo significado que 'haya' en el suyo, aun cuando refieran a árboles diferentes?

que determinan la referencia: la contribución de la sociedad y la contribución del mundo real. Una teoría semántica mejor tiene que abarcar a ambas.

TRADUCTORES: Eduardo Barrio, Patricia Brunsteins, Margarita Roulet y Julia Vergara.
REVISIÓN TÉCNICA: Eduardo Rabossi.

CAPÍTULO 9

EL INDIVIDUALISMO Y LA PSICOLOGÍA * (SELECCIÓN)

Tyler Burge **

Los años recientes han sido testigos del desarrollo de una semblanza de acuerdo en la psicología y en áreas coincidentes de la lingüística, la inteligencia artificial y las ciencias sociales, respecto de una manera de enfocar el estudio empírico de la capacidad y habilidad humanas. El enfoque es claramente mentalista ya que involucra la atribución de estados, procesos y eventos que son intencionales, en el sentido de "representacionales". Muchos de esos eventos y estados son inconcientes e inaccesibles a la mera reflexión. Generalmente se los etiqueta con jerga computacional. Pero se los puede comparar a los pensamientos, deseos, recuerdos, percepciones, planes, estados mentales o cosas parecidas, como se los llama de manera habitual. Tal como [ocurre] con las actitudes proposicionales corrientes, algunos son descriptos por medio de cláusulas subordinadas y pueden ser evaluados como verdaderos o falsos. Todos están involucrados en un sistema mediante el cual una persona conoce, representa y utiliza la información relativa a su alrededor [*surroundings*].

En [esta] parte del artículo criticaré algunos argumentos que se han dado para pensar que la explicación en psicología es y debe ser puramente "individualista"... Lo que tengo que decir a lo largo de este artículo tendrá relevancia para las partes de la psicología que atribuyen estados intencionales. Pero haré especial referencia a la explicación en la psicología cognitiva.

El individualismo [*individualism*] es un punto de vista acerca de [*about*] cómo se individuan correctamente las clases, [acerca] de cómo

* *The Philosophical Review* 95 (1986). Con autorización del autor y *The Philosophical Review*.

** Presenté una versión de este artículo en la conferencia Sloan en el MIT, en mayo de 1984. Me fueron útiles los comentarios de Ned Block, Fred Dretske y Stephen Stich. También aproveché la discusión con Jerry Fodor, David Israel, Bernie Kobes y Neil Stillings; estoy agradecido a los editores por varias sugerencias.

se fijan sus naturalezas. Nos ocuparemos principalmente del individualismo acerca de la individuación [*individuation*] de las clases mentales [*mental kinds*]. De acuerdo con el individualismo acerca de la mente, las naturalezas mentales de todos los estados (y eventos) mentales de una persona o de un animal, son tales que no hay una relación individuativa necesaria o profunda entre estar el individuo en esos tipos de estados y la naturaleza de su entorno [*environment*] físico y social.

La prominencia de este punto de vista se debe a Descartes. Fue adoptado por Locke, Leibniz y Hume. Recientemente ha logrado hacerse un lugar en la tradición fenomenológica y en las doctrinas de los conductistas, de los funcionalistas y de los teóricos de la identidad mente-cerebro del siglo veinte. Existen varias versiones más específicas de la doctrina. Algunos temas fundamentales de la filosofía tradicional están conformados por ella. En este artículo, sin embargo, me concentraré en las versiones de la doctrina que han tenido prominencia en la reciente filosofía de la psicología.

Los enfoques individualistas actuales de los estados y eventos mentales intencionales, han tendido a tomar una de las dos siguientes formas. Una sostiene que estar en un estado intencional dado (o ser el sujeto de tal evento) puede *ser explicado* por referencia a estados y eventos del individuo que son especificables sin usar un vocabulario intencional y sin presuponer nada acerca de su entorno social o físico. Se supone que la explicación especifica —en términos no intencionales— estímulos, conducta y estados internos físicos o funcionales del individuo. La otra forma de individualismo está implicada por la primera, pero es más débil. No intenta explicar nada. Simplemente hace una apelación a la *superveniencia* [*supervenience*]: los estados y eventos intencionales de un individuo (tipos y ejemplares [*tokens*]) no podrían ser diferentes de lo que son, dadas las historias físicas, químicas, neurales o funcionales del individuo; esas historias se especifican de manera no-intencional y de un modo que es independiente de las condiciones físicas o sociales externas al cuerpo del individuo.

En otros artículos he argumentado que ambas formas de individualismo están equivocadas. Los estados y eventos intencionales de una persona podrían variar (contrafácticamente), aun cuando permaneciera constante la historia física, funcional (y tal vez fenomenológica) del individuo, especificada de manera no intencional e individualista. He ofrecido varios argumentos a favor de esta conclusión. Apreciar la fuerza de esos argumentos y discernir el potencial filosófico de una perspectiva no individualista de la mente, depende en gran medida de reconsiderar las

diferencias entre esos argumentos. Ambos se refuerzan mutuamente y ayudan a relevar la topografía de una posición positiva.

Sin embargo, a los fines presentes, sólo esbozaré un par de argumentos para mostrar su tono característico. No los defenderé ni introduciré aclaraciones relevantes. Considérese una persona *A* que piensa que el aluminio es un metal liviano que se usa para los mástiles de los barcos y una persona *B* que cree que tiene artritis en el muslo. Supongamos que *A* y *B* pueden identificar [*pick out*] casos de aluminio y de artritis (respectivamente) y que conocen un buen número de hechos generales acerca del aluminio y la artritis. Sin embargo, *A* ignora la estructura química y las micropropiedades del aluminio. *B* ignora el hecho de que la artritis sólo puede darse en las articulaciones. Podemos imaginar casos contrafácticos en los cuales los cuerpos de *A* y de *B* tengan sus mismas historias consideradas ahora con prescindencia de sus entornos físicos, pero en las que existan diferencias ambientales significativas respecto de la situación real. El entorno contrafáctico de *A* carece de aluminio y tiene en cambio un metal liviano de aspecto similar. El entorno contrafáctico de *B* es tal que nadie aisló nunca a la artritis como una enfermedad específica o como un síndrome de enfermedades. En esos casos, *A* carecería de "pensamientos de aluminio" y *B* carecería de "pensamientos de artritis". Suponiendo patrones naturales de desarrollo, ambos tendrían diferentes pensamientos. Así, esas diferencias respecto de la situación real no sólo destacan las relaciones de los protagonistas con sus entornos, sino también con sus estados y eventos mentales intencionales, así llamados comúnmente. Los argumentos realzan las variaciones en las expresiones que ocurren oblicua o intencionalmente, en las adscripciones literales de estados y eventos mentales: nuestro principal instrumento para identificar estados mentales intencionales.[1]

1. "Individualism and the Mental", *Midwest Studies* 4 (1979), págs. 73-121; "Other Bodies", en *Thought and Object*, Woodfield (comp.) (Oxford, Oxford University Press, 1982); "Two Thought Experiments Reviewed", *Notre Dame Journal of Formal Logic* 23 (1982), págs. 284-293; "Cartesian Error and the Objectivity of Perception", en *Subject, Thought, and Context*, MacDowell y Pettit (comps.) (Oxford, Oxford University Press, 1986); "Intellectual Norms and Foundations of Mind" (de próxima aparición en *The Journal of Philosophy*). El argumento del aluminio es una adaptación del argumento de Hilary Putnam, "The Meaning of 'Meaning' ", *Philosophical Papers* Vol. II (Cambridge, Inglaterra, Cambridge University Press, 1975). Estrictamente, lo que Putnam escribió no fue compatible con este argumento. (Cf. el primero de los dos artículos citados en esta nota para la discusión). Pero el argumento del aluminio, en su superficie, está cerca del argumento que él da. El argumento de la artritis plantea dificultades a pesar de su paralelo metodológico.

Creo que estos argumentos usan descripciones literales de los eventos mentales y son independientes de los mecanismos conversacionales que pueden afectar la forma de una adscripción sin tener relación con la naturaleza del evento mental descripto. El tipo de argumento que hemos ilustrado no depende de los rasgos específicos de las nociones de artritis o de aluminio. Tales argumentos valen tanto para nociones observacionales como para nociones teóricas, para los perceptos como para los conceptos, para las nociones de clase-natural como para las de clase no-natural, para nociones que son propias de los expertos como para las nociones que la literatura psicológica conoce como "categorías básicas" ["*basic categories*"]. Pienso por cierto que como mínimo se pueden formular argumentos similares relevantes que valgan para tipos públicos de objetos, propiedades o eventos, conocidos, de manera típica a través de medios empíricos.[2]

No elaboraré ni defenderé aquí estos argumentos. En lo que sigue presupondré que son sólidos. Para nuestros propósitos será suficiente tener en mente su conclusión: los estados y eventos mentales pueden variar, en principio, junto [*with*] con cambios en el entorno, aun si la historia física (funcional, fenomenológica), especificada de manera no intencional e individualista, permanece constante.

Una reacción común a estas conclusiones, que a menudo no es apoyada en argumentos, ha consistido en conceder su fuerza pero limitar su efecto. Se sostiene con frecuencia que se aplican a las atribuciones de actitudes de sentido común, pero que no son aplicables a atribuciones análogas en la psicología. Se ha sostenido que los aspectos no individualistas de una atribución mentalista no congenian con los propósitos y requerimientos de una teoría psicológica. Por supuesto, hay toda una tradición que sostiene que las atribuciones intencionales corrientes son incapaces de producir conocimiento alguno. Otros han sostenido el punto de vista más modesto de que las atribuciones mentalistas sólo son capaces de producir un conocimiento que, en principio, no podría ser sistematizado en una teoría.

No podré discutir todas esas líneas de pensamiento. En particular, ignoraré los argumentos generales de que las adscripciones mentalistas son profundamente indeterminadas o incapaces de producir conoci-

2. Sobre las categorías básicas, cf., *e.g.*, Rosch, Mervis, Gray, Johnson, Boyes-Graem, "Basic Objects in Natural Categories", *Cognitive Psychology* 8 (1976), págs. 382-489. Sobre la afirmación general en esta última oración, cf. "Intelectual Norms", *op. cit.* y la última parte de este artículo.

miento. Nuestra atención se centrará en los argumentos que se proponen mostrar que las adscripciones mentalistas no-individualistas no pueden desempeñar un rol sistemático en la explicación psicológica, *debido* al hecho de que no son individualistas.

Hay por cierto diferencias significativas entre el discurso teorético en la psicología y el discurso mentalista de sentido común. La más evidente es que el lenguaje de la psicología teorética requiere refinar el discurso ordinario. No solamente requiere mayor sistematicidad y rigor, y un cúmulo de estados y de eventos inconcientes que corrientemente no son atribuidos (aunque son, pienso, comúnmente admitidos). También tiene que distinguir los propósitos descriptivo-explicativos de las atribuciones corrientes, de los usos que sirven a la comunicación a expensas de la descripción y la explicación. Hacer esta distinción es casi una práctica común. El refinamiento con fines científicos tiene que ser, sin embargo, sistemático y meticuloso, aunque no necesita eliminar toda la vaguedad. Pienso que no hay razones de peso para creer que un refinamiento tal no pueda ser logrado a través del desarrollo de una teoría psicológica, o que al efectuarlo cambiará fundamentalmente la naturaleza de las atribuciones mentalistas corrientes.

Las diferencias entre el discurso ordinario y el científico sobreviven aun cuando el discurso ordinario experimente los refinamientos recién mencionados. Si bien el discurso de sentido común —tanto respecto de los objetos macrofísicos como de los eventos mentales— produce conocimiento, creo que los principios que gobiernan la justificación de tal discurso difieren de los que se invocan en el teorizar científico sistemático. De este modo hay espacio, prima facie, para el punto de vista de que la psicología es o debería ser totalmente individualista, aun cuando las descripciones corrientes de los estados mentales no lo sean. Sin embargo, los argumentos que se han ofrecido a favor de este punto de vista, no me parecen sólidos. Tampoco encuentro que este punto de vista sea persuasivo.

Antes de considerar tales argumentos, tengo que articular algunas suposiciones adicionales de fondo, esta vez respecto de la psicología misma. Voy a tomar a las partes de la psicología que utilizan un discurso mentalista y de procesamiento de la información, aproximadamente como son. Doy por supuesto que emplean la metodología científica estándar, que han producido interesantes resultados empíricos y que contienen algo más que una muestra de teoría genuina. No voy a prejuzgar qué tipo de ciencia es la psicología, o cómo se relaciona con las ciencias naturales. No obstante, voy a suponer que sus alegaciones cog-

nitivas y, más especialmente, sus métodos y presuposiciones deben ser tomados seriamente como lo mejor que tenemos por ahora en esta área de la investigación. Creo que no hay buenas razones para pensar que los métodos o hallazgos de este cuerpo de investigación estén radicalmente equivocados.

No supondré que la psicología *tenga que* continuar manteniendo contacto con el discurso de sentido común. Creo que tal contacto se va a mantener, casi seguramente. No obstante creo que las disciplinas empíricas deben encontrar su propio camino de acuerdo con los estándares que se plantean a sí mismas. Los reparos casi a priori formulados por los filósofos, cuentan poco. De modo que nuestras reflexiones tienen que ver con la psicología tal como es, no como será o tendría que ser.

Al tomar a la psicología tal como es, estoy suponiendo que busca refinar, profundizar, generalizar y sistematizar algunos de los enunciados propios de un sentido común ilustrado [*informed common sense*] acerca de la actividad mental de la gente; que acepta, por ejemplo, que la gente ve objetos físicos con ciertas configuraciones [*shapes*], texturas y matices, y en ciertas relaciones espaciales, bajo ciertas condiciones especificadas, e intenta explicar con mayor profundidad lo que la gente hace cuando ve tales cosas, y cómo su hacer es realizado. La psicología acepta que la gente recuerda eventos y verdades, que categoriza objetos, que efectúa inferencias, que actúa de acuerdo con creencias y preferencias. [La psicología] intenta encontrar regularidades profundas en esas actividades, especificar los mecanismos que subyacen a ellas y proveer explicaciones sistemáticas de cómo esas actividades se relacionan entre sí. Al describir y, al menos en parte, explicar esas actividades y habilidades, la psicología hace uso de cláusulas subordinadas [*that-clauses*] interpretadas y de otras construcciones intencionales; de lo que podríamos llamar en forma laxa "contenido intencional" ["*intentional content*"].[3] No he visto ninguna razón de peso para creer que este uso sea meramente heurístico, instrumental o de segunda clase, en algún otro sentido.

Doy por supuesto que el contenido intencional tiene estructura

3. Nuestro hablar de "contenido" intencional será ontológicamente descolorido. Puede ser convertido en un hablar acerca de cómo las cláusulas subordinadas (o sus componentes) son interpretadas y diferenciadas —tomadas como equivalentes o no equivalentes— para los propósitos cognitivos de la psicología. No todos los estados o estructuras intencionales que son atribuidos en psicología son explícitamente proposicionales. Mi punto de vista en este artículo vale, en general, para estados intencionales.

interna —algo parecido a una estructura gramatical o lógica— y que las partes de esta estructura se individuan de modo suficientemente fino [*finely enough*] como para corresponder a ciertas habilidades, procedimientos o perspectivas individuales. Ya que distintas habilidades, procedimientos o perspectivas pueden ser asociados con cualquier evento, objeto, propiedad o relación dados, el contenido intencional tiene que ser individuado más finamente que las entidades mundanas con las que el individuo interactúa. Tenemos que permitir los diferentes modos (aun, pienso, diferentes modos primitivos) en los que el individuo concibe o representa una entidad dada. El supuesto de que en la psicología el contenido sea de grano fino, no jugará un rol explícito en lo que sigue. Lo menciono aquí para indicar que mi escepticismo acerca del individualismo como una interpretación de la psicología, no es el resultado de una concepción del contenido de la que resulta claro ya que no cumple un papel dominante en la psicología.[4]

Finalmente, supondré que el individualismo está prima facie *equivocado*, respecto de la psicología, incluyendo a la psicología cognitiva. Puesto que las partes relevantes de la psicología utilizan frecuentemente atribuciones de estados intencionales que son sometidos a nuestros experimentos mentales [*thought experiments*], el lenguaje que en realidad se usa en la psicología no es puramente individualista. Es decir, las generalizaciones con fuerza contrafáctica que aparecen en las teorías psicológicas, no son todas individualistas en sus interpretaciones estándar. Para la comprensión corriente de las condiciones de verdad o para las condiciones de individuación de las atribuciones relevantes, basta con verificar los experimentos mentales. Más aún, no existe en la actualidad un lenguaje individualista bien explicado, bien comprendido y mucho menos bien testeado, o una reinterpretación individualista de las formas lingüísticas que corrientemente se usan en psicología, que pudiera servir de sustituto.

4. Ciertos enfoques en la lógica intensional que dan preferencia a la "referencia directa" o bien a alguna analogía entre las actitudes y la necesidad, han argumentado que esta manera de estructurar finamente el contenido de una actitud, debe ser revisado. Pienso que por razones puramente filosóficas esos enfoques no pueden dar cuenta de las actitudes. Por ejemplo, sirven poco para iluminar las numerosas variaciones de la "paradoja de la identidad" de Frege. Parecen ser aún menos recomendables como prescripciones para el lenguaje de la psicología. Algunas defensas del individualismo consideran que estos enfoques acerca del contenido proposicional constituyen una oposición al individualismo. Pienso que esos enfoques no son serios rivales como explicación de las actitudes proposicionales y que por eso deben ser dejados fuera de la discusión.

De este modo, el individualismo, tal como se aplica a la psicología, tiene que ser revisionista. Tiene que serlo al menos respecto del lenguaje de la teoría psicológica. Desarrollaré el punto de vista de que es también revisionista, sin una buena razón, respecto de las presuposiciones subyacentes de la ciencia. Para justificarse a sí mismo, el individualismo tiene que cumplir con dos tareas. Tiene que mostrar que el lenguaje de la psicología debería ser revisado, demostrando que las presuposiciones de la ciencia son o deberían ser *puramente* individualistas. Y tiene que explicar un lenguaje individualista nuevo (que atribuya lo que a veces se llama "contenido estrecho" ["*narrow content*"]) que capte los compromisos teóricos genuinos de la ciencia.

Estas tareas son independientes. Si se cumpliera la segunda y permaneciera sin cumplirse la primera, el individualismo estaría equivocado; pero habría generado un nuevo nivel de explicación. Por razones que luego mencionaré, soy escéptico acerca de tal complementación en masa de la teoría actual. Pero la psicología no es un todo monolítico. Coexisten dentro de ella diferentes tareas explicativas y tipos de explicación. Al cuestionar el punto de vista de que la psicología es individualista, no estoy *con ello* dudando de que haya algunas subpartes de la psicología que se ajusten a los reparos [*strictures*] del individualismo. Dudo de que toda la psicología, tal como es practicada actualmente, sea o deba ser individualista. Por eso intentaré satisfacer la primera de las dos tareas con las que se enfrenta alguien que está resuelto a revisar a la psicología de acuerdo con los lineamientos individualistas. Hasta aquí los comentarios preliminares.

I

Comenzaremos discutiendo un argumento muy general en contra de las explicaciones no-individualistas. El argumento es el siguiente. La conducta de los protagonistas de nuestros experimentos mentales, fisiológica y funcionalmente idénticos, es también idéntica. Pero la psicología es la ciencia (sólo) de la conducta. Puesto que la conducta de los protagonistas es la misma, una ciencia de la conducta debería dar las *mismas* explicaciones y descripciones de los dos casos (por algún principio ockamista de parsimonia). Por lo tanto, no hay espacio en la disciplina para explicar sus conductas en términos de estados mentales diferentes.[5]

5. Stephen Stich, *From Folk Psychology to Cognitive Science* (Cambridge, Mass.,

Las dos premisas iniciales son problemáticas. Comencemos por la primera: no tiene que suponerse que los protagonistas de los experimentos mentales sean conductualmente idénticos. Creo que la única interpretación clara y general de "conducta" de la que disponemos y que verificaría la primera premisa, es "movimiento corporal" ["*bodily motion*"]. Pero esta interpretación casi no tiene relevancia en la psicología tal como se la practica actualmente. En la psicología, "conducta" se ha convertido en un término abarcativo para la actividad observable respecto de cuya descripción y carácter los psicólogos pueden alcanzar un rápido acuerdo "preteórico". Dejando a un lado prejuicios metodológicos, no es verdad que todas las descripciones que podrían contar como "conductuales" en la psicología cognitiva (social, evolutiva), se aplicarían a los protagonistas. La mayor parte de la conducta es acción intencional; muchas especificaciones de la acción son no-individualistas. Experimentos mentales similares, de manera relevante, a los que ya hemos desarrollado, valdrían para ellas.

Por ejemplo, gran parte de la evidencia "conductual" se obtiene en la psicología de lo que la gente dice o de cómo contesta preguntas. Las proferencias [*utterances*] de los sujetos (y las preguntas que se les formulan) tienen que ser consideradas como interpretadas a fin de que sean de algún uso en los experimentos, y a menudo se supone que las teorías pueden ser controladas por experimentos que se llevan a cabo en lenguajes diferentes. Dado que los decires [*sayings*] de los protagonistas en los experimentos mentales son diferentes, aun en los casos no transparentes u oblicuos, es prima facie erróneo considerar a los protagonistas como "conductualmente" idénticos. Muchas atribuciones de la conducta no verbal son también intencionales y no-individualistas, o aun relacionales: ella recoge una manzana, señala el edificio, sigue la trayectoria de la pelota, sonríe a un rostro familiar, toma el dinero en vez de correr el riesgo. Estas atribuciones pueden elaborarse para producir experimentos mentales no individualistas. El punto general a señalar es que en psicología muchas especificaciones relevantes de la conducta son intencionales o relacionales, o ambas. Los experimentos mentales indican que estas especificaciones fundamentan atribuciones mentales no-individualistas.

MIT Press, 1983), capítulo 8 [incluido en esta compilación, págs. 205-26]. Aunque no discutiré el principio ockamista no formulado, soy escéptico respecto de él. Aparte de la circularidad de sus supuestos, no encuentro claro por qué debería exigirse a una ciencia que explique de la misma manera a dos instancias del mismo fenómeno, en especial si las condiciones del entorno conducen a una diferencia en tales instancias.

Un argumento a favor del individualismo no puede *suponer* razonablemente que esas especificaciones sean individualistas o deban serlo.

Hay, por supuesto, especificaciones no-individualistas de la conducta que son inadecuadas para cualquier empresa científica ("el movimiento corporal favorito de mi amigo"). Pero la mayoría de ellas no aparecen siquiera en la psicología. El problema de proporcionar especificaciones razonables de la conducta no puede resolverse desde un sillón. Sanear la noción de conducta para que satisfaga algún principio metodológico sostenido con anterioridad, es un viejo juego que nunca se ha ganado. Uno tiene que atender a lo que la psicología realmente toma como evidencia "conductual". Es responsabilidad del argumento mostrar que las nociones no-individualistas no tienen cabida en la psicología. En tanto el argumento suponga que las especificaciones intencionales, no-individualistas de la conducta son ilegítimas, ignora aspectos obvios de la práctica psicológica o bien presupone la cuestión en discusión.

El segundo paso del argumento también flaquea. No se puede suponer sin una discusión seria que la psicología es caracterizada correctamente como una ciencia (sólo) de la conducta. Por supuesto que es así, si la conducta es interpretada de un modo restrictivo. Pero, aun dejando a un lado cómo la conducta sea interpretada, la premisa es dudosa. Una razón es que difícilmente se tenga que suponer que una ciencia putativa haya de ser caracterizada en términos de su evidencia, como opuesta a su tema específico. Por supuesto, el tema específico está en alguna medida en discusión. Pero la psicología cognitiva parece ocuparse de ciertas habilidades y actividades molares [*molar*], algunas de las cuales son las actitudes proposicionales. Puesto que las actitudes proposicionales atribuidas no parecen ser completamente individuables en términos individualistas, necesitamos un argumento directo a favor de que la psicología cognitiva no es una ciencia de aquello de lo que parece serlo.

Una segunda razón para poner en duda la premisa es que la psicología parece ocuparse en parte de las relaciones entre la gente o los animales y su entorno. Es difícil encontrar el modo de proveer una descripción natural de una teoría de la visión, por ejemplo, como una ciencia de la conducta. El punto central de la teoría es comprender cómo la gente hace lo que hace, obviamente con éxito, cómo ve objetos en su entorno. Tratamos de explicar las relaciones entre un sujeto y un mundo físico, del cual consideramos saber algo. Las teorías de la memoria, de ciertos tipos de aprendizaje, de comprensión lingüística, de formación de creencias, de categorización, hacen lo mismo. No es obvio, cierta-

mente, que estas referencias a las relaciones entre el sujeto y su entorno sean de alguna manera inesenciales a (todas las áreas de) una teoría psicológica. Parecen ser, de hecho, una parte extensa del objetivo central de esa teoría. En mi opinión, esas relaciones ayudan a motivar principios de individuación no individualistas... En suma, pienso que el argumento que hemos considerado hasta aquí es circular en casi todos sus pasos significativos.

Existe un argumento emparentado que vale la pena considerar: los determinantes [*determinants*] de la conducta supervienen a los estados del cerebro. (Si uno es un materialista, podría considerar lo que sigue como una trivialidad: "Los estados cerebrales supervienen a los estados cerebrales"). Por eso, si las actitudes proposicionales han de ser consideradas entre los determinantes de la conducta, tienen que ser tratadas como superviniendo a los estados cerebrales. La alternativa consiste en considerar a las actitudes proposicionales como conductualmente irrelevantes.[6]

Pienso que este argumento puede darse vuelta. Puesto que las actitudes proposicionales cuentan como determinantes de nuestra "conducta" (donde la expresión es tan vaga como siempre), y puesto que las actitudes proposicionales no supervienen a nuestros estados cerebrales, no todos los determinantes de nuestra "conducta" supervienen a nuestros estados cerebrales. Deseo formular tres puntos centrales en contra del argumento original, dos metafísicos y uno cognoscitivo [*epistemic*] o metodológico. Primero la metafísica.

La apuesta ontológica que supone la doctrina de la superveniencia es mucho menos substancial de lo que uno podría pensar. No es una "consecuencia trivial" del materialismo de los estados y eventos mentales el que los determinantes de nuestra conducta supervengan a los estados de nuestros cerebros. Esto es así porque qué superviene a qué tiene, al menos, tanto que ver con cómo son individuadas las entidades

6. No pude encontrar una enunciación completamente explícita de este argumento en algún trabajo publicado. Parece informar, algunos pasajes de Jerry Fodor, "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology" en *Representations* (Cambridge, Mass., MIT Press, 1981), *e.g.*, págs. 228-232. Está cerca de la superficie en muchos trabajos influidos por los artículos de Fodor. Cf., Colin McGinn, "The Structure of Content" en Woodfield (comp.), *Thought and Object* (Oxford, Clarendon Press, 1982), págs. 207-216. Quienes como McGinn conceden la fuerza de los argumentos en contra del individualismo, utilizan algo parecido a este argumento para afirmar que los "aspectos" individualistas de los estados intencionales son los que son relevantes a la explicación psicológica.

relevantes, como con aquello de lo que están hechas. Si un evento mental *m* es individuado en parte por referencia a condiciones normales externas al cuerpo de una persona, entonces, sin que importe si *m* tiene una composición material, *m* podría variar aun cuando el cuerpo permaneciera siendo el mismo.

Ya que los fenómenos intencionales constituyen un caso especial tan extenso, es probablemente engañoso buscar analogías en otros dominios para ilustrar el punto. Sin embargo, para liberar la imaginación, consideremos la batalla de Hastings. Supongamos que preservamos cada cuerpo humano, cada trozo de césped, cada arma, cada estructura física y todas las interacciones físicas entre ellos, desde la primera confrontación hasta la última muerte o retirada el día de la batalla. Supongamos que, contrafácticamente, imagináramos todos estos eventos y sus accesorios físicos localizándolos en California (quizás en el mismo momento, en 1066). Supongamos que la actividad física sea inducida artificialmente por científicos brillantes transportados a la tierra por productores marcianos de cine. Las causas lejanas [*distal*] de la batalla no tienen nada que ver con las causas de la batalla de Hastings. Pienso que es plausible (y ciertamente coherente) decir que en tales circunstancias no habría tenido lugar la Batalla de Hastings, sino solamente un facsímil físico. Pienso que aun si se respetara el lugar en que acaeció la Batalla de Hastings, antecedentes causales contrafácticos suficientemente diferenciados alcanzarían para variar la identidad de la batalla. La batalla es individuada, en parte, en términos de sus causas. Aunque la batalla no superviene a sus constituyentes físicos, casi ni dudamos en considerarla un evento físico.

Nuestra individuación de batallas históricas está probablemente relacionada con los estados intencionales de los participantes. Este punto también se puede hacer con referencia a los casos que son claramente independientes de consideraciones intencionales. Pensemos en la emergencia oceánica de América del Norte. Supongamos que delimitamos qué cuenta como eventos físicos (digamos, microfísicos) constituyentes de ese evento. Parece que si las condiciones físicas circundantes y las leyes son ideadas con ingenio, podemos concebir contrafácticamente esos mismos eventos constituyentes (o los objetos físicos constituyentes que en los mismos lugares padecen cambios físicamente idénticos) de manera tal de incluirlos dentro de una masa territorial más grande, de modo que los constituyentes físicos de América del Norte no formen ninguna parte saliente de esta masa mayor. En tal caso, la emergencia de América del Norte no habría ocurrido, aun si sus eventos físicos

"constituyentes" fueran, de manera aislada, físicamente idénticos a los eventos reales. Individuamos la emergencia de los continentes u otras masas territoriales de modo tal que no supervengan a sus constituyentes físicos. No obstante, tales eventos son físicos.

Pienso, en realidad, que el materialismo no provee restricciones razonables a las teorías respecto del rol de las atribuciones mentalistas en psicología. La relación de composición física no juega en realidad un rol significativo en ninguna teoría científica sólida de los eventos mentales o de sus relaciones con eventos cerebrales. Las restricciones que las consideraciones fisiológicas imponen a la teorización psicológica son, aunque substanciales, más débiles que las de cualquiera de los materialismos que se hayan elaborado; aun la variedad composicional débil a la que estoy aludiendo. Mi punto consiste precisamente en que refutar la superveniencia individualista no implica rechazar un punto de vista materialista. Así, el materialismo per se no hace nada que cuente como apoyo al individualismo.[7]

El segundo punto "metafísico" concierne a la causación. El argumento que estamos considerando en realidad supone simplemente que las actitudes proposicionales (tipo y ejemplar) supervienen a eventos físico-químicos en el cuerpo. Sin embargo, muchos filósofos parecen pensar que esta suposición se torna obvia a partir de ciertas observaciones tibias sobre la etiología de las conductas y de los eventos mentales. Es plausible que los eventos en el mundo externo afecten causalmente a los eventos mentales de un sujeto afectando solamente a las superficies corporales de dicho sujeto; y que nada (sin excluir a los eventos

7. En "Individualism and the Mental", *op. cit.*, págs. 109-113, argumento que las teorías de la *identidad* de ejemplares (casos) son consideradas implausibles debido a los experimentos mentales no-individualistas. Pero, las teorías de la identidad de ejemplares no son el último bastión para una estrategia de defensa materialista. Lo que es crucial es la composición. Es coherente, aunque pienso que es erróneo, sostener que las atribuciones de actitudes proposicionales no atrapan rígidamente a los eventos físicos: así las actitudes proposicionales varían entre los protagonistas actuales y contrafácticos en los experimentos mentales, aunque la ontología de los eventos mentales ejemplares continúe siendo idéntica. Este punto de vista es compatible con gran parte de mi oposición crítica al individualismo. Pero pienso que no hay una buena razón para creer la tesis implausible de que los eventos mentales no son individuados ("esencialmente" o "básicamente") en términos de las atribuciones relevantes de actitudes proposicionales (cf. *ibid.*). Así rechazo el punto de vista de que los mismos eventos mentales (ejemplares o tipos) sean referidos por diferentes descripciones en los experimentos mentales. Estas consideraciones subyacen mi recomendación al materialista convencido, de que la composición más que la identidad es el paradigma. (Yo sigo sin convencerme.)

mentales) afecte causalmente su conducta salvo afectando a (causando o siendo un antecedente causal de causas de) los estados locales del cuerpo del sujeto. Se podría especular que en los experimentos mentales anti-individualistas esos principios son violados en tanto que los eventos en el entorno son alegados para "afectar" los eventos mentales y la conducta de una persona, sin "afectar" de manera diferencial su cuerpo: sólo si los eventos (y estados) mentales supervienen a los cuerpos de los individuos, los principios causales pueden ser mantenidos.

El razonamiento es confuso. La confusión es instigada por un uso no cuidadoso del término "afectar" ["*affect*"], que confunde la causación con la individuación. Las variaciones en el entorno que no varían los impactos que "afectan" causalmente al cuerpo del sujeto, pueden "afectar" la individuación de la información que el sujeto está recibiendo, o de los procesos intencionales que él o ella está protagonizando, o la manera en que el sujeto está actuando. No se sigue que el entorno afecte causalmente al sujeto de modo tal que le impida tener efectos en el cuerpo del sujeto.

Una vez que se evita la confusión, resulta claro que no hay un argumento simple a favor del individualismo, que parta de los principios causales enunciados. El ejemplo de la geología brinda un modelo opuesto útil. Muestra que uno puede aceptar los principios causales y con ello no experimentar ninguna perplejidad al rechazar el individualismo. Un continente se mueve y es movido por rocas, olas, moléculas. No obstante, podemos concebir mantener constantes los impactos periféricos y los eventos y objetos químicamente constituyentes del continente, sin mantener idéntico el continente o algunos de sus macrocambios, debido a que las relaciones espaciales del continente con otras masas territoriales afectan el modo como nosotros lo individuamos. Tomemos un ejemplo de la biología. Aceptemos el principio plausible de que nada afecta causalmente la respiración a menos que afecte causalmente los estados locales de los pulmones. No se sigue, y realmente no es verdad, que individuemos los pulmones y los variados subeventos-eventos de la respiración de modo tal que tratemos esos objetos y eventos como supervenientes a los objetos y eventos componentes descriptos químicamente. Si el mismo proceso químico (el mismo a partir de la superficie de los pulmones, dentro y detrás de las superficies) fuera incluido en un tipo de cuerpo diferente y tuviera una función completamente distinta (digamos, digestiva, inmunológica o regulatoria), no estaríamos tratando con los mismos estados y eventos biológicos. La causación local no hace más plausible la individuación local o la superveniencia individualista.

La analogía con los eventos mentales debería ser evidente. Podemos concordar en que la conducta y los eventos mentales de una persona son afectados causalmente por su entorno solamente a través de los efectos causales locales en su cuerpo. Sin la más leve disconformidad conceptual, podemos individuar eventos mentales, de modo tal que permitan diferenciar los eventos (tipos o ejemplares) con químicas o aun con fisiologías indistinguibles para el cuerpo del sujeto. La información del entorno sólo se transmite a través de las estimulaciones próximas [*proximal*], pero la información es en parte individuada por referencia a la naturaleza de los estímulos normales lejanos [*distal*]. La causación es local. La individuación puede presuponer hechos acerca de la naturaleza específica del entorno de un sujeto.

En tanto que la explicación psicológica intencional es en sí misma causal, bien puede presuponer que las transacciones causales a las que se aplican sus generalizaciones comparten alguna relación necesaria con ciertas transacciones físicas subyacentes (u otras). Sin un conjunto de transacciones físicas, ninguna de las transacciones intencionales ocurriría. Pero no se sigue que las clases invocadas en la explicación de las interacciones causales entre estados intencionales (o entre estados físicos y estados intencionales, por ejemplo, en la visión o en la acción) supervengan a las transacciones fisiológicas subyacentes. Las mismas transacciones físicas en una persona pueden en principio mediar o subyacer a transacciones que involucran distintos estados intencionales, si los rasgos del entorno que entran en la individuación de estados intencionales y que son críticos en las generalizaciones explicativas que invocan aquellos estados, varían de las maneras apropiadas.

Vayamos a nuestro argumento cognoscitivo. El punto de vista de que las actitudes proposicionales ayudan a determinar la conducta está bien arraigado en los juicios corrientes y en las prácticas explicativas de la psicología. Nuestros argumentos de que las actitudes proposicionales de un sujeto no están fijadas solamente por sus estados cerebrales se basan en juicios ampliamente compartidos acerca de casos *particulares* que en los aspectos relevantes muestran elementos familiares en nuestras prácticas efectivas psicológicas y de sentido común de atribución de actitudes. Por contraposición, la afirmación de que ninguna de las actitudes proposicionales de un individuo (o los determinantes de su conducta) podría haber sido diferente a menos que alguno de sus estados cerebrales fuera distinto, es una conjetura metafísica. Es una generalización modal que no está fundamentada en juicios acerca de casos particulares o (hasta ahora) en una cuidadosa interpretación de la explica-

ción real y de las prácticas descriptivas de la psicología. La ideología metafísica debería o bien estar de acuerdo con la praxis intelectual e iluminarla, o bien producir fuertes razones para revisarla.

Lo que sabemos acerca de la superveniencia tiene que ser derivado, en parte, de lo que sabemos acerca de la individuación. Lo que sabemos de la individuación se deriva de la reflexión acerca de las explicaciones y descripciones de las prácticas cognitivas actuales. Los métodos de individuación están estrechamente ligados a las necesidades explicativas y descriptivas de tales prácticas. De este modo, los juicios justificados respecto de qué superviene a qué son *derivados* de la reflexión sobre la naturaleza de la explicación y descripción en el discurso psicológico y de las atribuciones de actitud comunes. Pienso que tales juicios no pueden ser razonablemente invocados para restringir tal discurso. Me parece en consecuencia que sin entrar en un argumento adicional, la tesis de la superveniencia individualista no brinda razón para reclamar el (pan-)individualismo en psicología. De hecho, la argumentación a partir de la superveniencia individualista es una petición de principios. *Presupone* más que establece que la *individuación* en psicología —*con ello la explicación y descripción*— debería ser completamente individualista. Es, simplemente, el tipo erróneo de consideración a invocar en una discusión acerca de la explicación y la descripción.

Pienso que esta observación es bastante general. No son sólo los problemas de la superveniencia, sino también los problemas de ontología, reducción y causación en general los que son cognoscitivamente posteriores a las cuestiones relativas al éxito de las prácticas explicativas y descriptivas.[8] No puede criticarse razonablemente una pretendida práctica explicativa o descriptiva apelando principalmente a alguna concepción previa sobre qué sea "una buena entidad", o sobre cómo debería ser la individuación o la referencia, o sobre qué aspecto debería tener la estructura total de la ciencia (o del conocimiento). Las cuestio-

8. Los puntos sobre ontología y referencia se remontan a Frege, *Foundations of Arithmetic*, Trad. de Austin (Northwestern University Press, Evanston, 1968). El punto acerca de la reducción es relativamente obvio, aunque unos pocos filósofos han propuesto concepciones acerca de la unidad de la ciencia con un espíritu relativamente apriorístico. Aplicado a la ontología, el punto es, al menos, básico para el pragmatismo de Quine. Me parece que hay, sin embargo, trazos en el trabajo de Quine y en el de la mayoría de sus seguidores que dejan que una preocupación por el fisicalismo afecte el "*insight*" pragmático de Frege (y Quine). Es simplemente ilusorio pensar que las preconcepciones metafísicas o aun cognoscitivas proveen un estándar para juzgar las ontologías o los esfuerzos explicativos de las ciencias particulares, deductivas o inductivas.

nes acerca de lo que existe, de cómo se individuan las cosas y de qué se reduce a qué, surgen en referencia a las prácticas descriptivas y explicativas corrientes. Las respuestas que se proponen a estos problemas no pueden utilizarse, por sí mismas, para criticar un modo, por otra parte exitoso, de descripción y explicación.[9]

Por supuesto, uno podría proponerse basar el principio individualista de superveniencia en lo que sabemos acerca de una buena explicación. Quizás, uno podría intentar argüir, partiendo de una inferencia a la mejor explicación [*from inference to the best explanation*] concerniente a las relaciones de más alto nivel, hasta llegar a teorías más básicas en las ciencias naturales que las entidades postuladas por la psicología como debiendo supervenir a las de la fisiología. O quizás, uno podría intentar trazar analogías entre teorías no individualistas en psicología y teorías pasadas no exitosas. Estas dos estrategias podrían satisfacer nuestros reparos metodológicos al responder la pregunta de si las explicaciones no-individualistas son viables de una manera en que una apelación pura a un principio de superveniencia no lo es. Pero las invocaciones filosóficas de la inferencia a la mejor explicación tienden a ocultar saltos bruscos apoyados principalmente por la ideología. Tales consideraciones tienen que ser expuestas por medio de argumentos. Hasta aquí no parecen ser muy prometedoras.

Consideremos la primera estrategia. Las inducciones que van de las ciencias naturales a las ciencias humanas son problemáticas desde el comienzo. Los problemas de estos dos tipos de ciencias parecen ser muy diferentes y de las más diversas maneras. Por supuesto, se puede intentar explotar razonablemente las analogías con un espíritu pragmático. Pero el hecho de que alguna analogía dada no se sostenga, difícilmente cuente en contra de un modo, por lo demás, viable de explicación. Más aún, hay modos de explicación no individualista aun en las ciencias naturales. La geología, la fisiología y partes de la biología apelan a entidades

9. En términos más generales aun, pienso que en la filosofía, el poder cognoscitivo deriva en gran medida de las reflexiones sobre las implementaciones particulares de las prácticas cognitivas exitosas. Por práctica cognitiva entiendo una empresa cognitiva que sea estable, que esté de acuerdo con las condiciones estándares de comprobación intersubjetiva y que incorpore un núcleo substancial de acuerdo entre sus practicantes. Por supuesto, las hipótesis filosóficas revisionistas no tienen que ser rechazadas sin más. Algunas veces, pero raramente en nuestros días, tales hipótesis ejercen influencia en las prácticas cognitivas, expandiendo la imaginación teórica para conducir a nuevos descubrimientos. Un cambio en la práctica puede vindicar las hipótesis filosóficas. Pero las hipótesis aguardan tal vindicación.

que no son supervenientes en su composición física subyacente. En estas ciencias las nociones de clase (placas, órganos, especies) presuponen métodos individuativos que hacen referencia esencial al entorno que rodea las instancias de tales clases.

La segunda estrategia parece aun menos prometedora. Tal como se plantea, se encuentra seriamente afectada por la vaguedad. Algunos autores han sugerido similitudes entre el vitalismo en la biología o las teorías de la acción a distancia en la física, y las teorías no individualistas en psicología. Las analogías son tenues. A diferencia del vitalismo, la psicología no individualista no apela ipso facto a un nuevo tipo de fuerza. A diferencia de las teorías de la acción a distancia, no apela a una acción a distancia. Es verdad que los aspectos del entorno que no afectan de manera diferencial al movimiento físico de los protagonistas en los experimentos mentales, sí afectan de manera diferencial las explicaciones y las descripciones. Sin embargo, esto no se debe a que se haya postulado alguna relación causal espacial sino, más bien, a que las diferencias del entorno afectan las clases de leyes que se obtienen y el modo como las causas y los efectos son individuados.

Consideremos ahora un tipo adicional de objeción a la aplicación de los experimentos mentales en la psicología. Dado que los protagonistas reales y contrafácticos son tan impresionantemente *similares* de tantas maneras psicológicamente relevantes, ¿puede un lenguaje teórico que se hace cargo de [*cuts across*] estas similitudes ser empíricamente adecuado? Las similitudes "conductuales" fisiológicas y no intencionales entre los protagonistas parecen demandar una similitud de explicación. En su forma más fuerte, esta objeción quiere señalar que no hay espacio en la psicología para un lenguaje mentalista no individualista. En su forma más débil, intenta motivar un nuevo lenguaje teorético que atribuya contenido intencional, aunque sea individualista. Sólo la forma más fuerte establecería el individualismo en psicología. La consideraré primero.

La objeción es que las similitudes entre los protagonistas torna implausible cualquier teoría que los trate de manera diferente. Esta objeción es vaga o entimemática. Precisarla tiende a retrotraernos a uno de los argumentos que ya hemos refutado. Cualquiera que sea el punto de vista, hay varios medios disponibles (la neurofisiología, partes de la psicología) para explicar de una manera parecida las similitudes que se postulan entre los protagonistas en los experimentos mentales. El argumento no tiene siquiera la forma correcta de producir una razón para pensar que las diferencias entre los protagonistas no debieran reflejarse

en algún lugar en la teoría psicológica; precisamente es este el punto en cuestión.

A menudo, la objeción se asocia con la observación de que las explicaciones no individualistas tornarían "milagrosos" los paralelos en la conducta de los protagonistas en los experimentos mentales: explicar los mismos fenómenos de conducta como resultantes de actitudes proposicionales diferentes sería invocar un "milagro". La retórica acerca de los milagros puede ser conjurada al advertir que la "conducta" de los protagonistas no es lisa y llanamente idéntica, que las explicaciones no individualistas postulan fuerzas no especiales, y que existen diferencias físicas en los entornos de los protagonistas que ayudan a motivar la descripción y la explicación de su actividad, de maneras diferentes, al menos en un nivel.

La retórica acerca de los milagros raya el límite de un malentendido fundamental acerca del *status* de los experimentos mentales no individualistas y de la relación entre la filosofía y la psicología. Por supuesto que hay una implausibilidad empírica considerable, que podríamos llamar con algo de exageración "milagrosa", en que dos personas tengan idénticas historias individualistas físicas pero pensamientos diferentes. Gran parte de esta implausibilidad es la artificialidad de la versión de "dos personas" de los experimentos mentales, un rasgo que es bastante inesencial. (Se puede tomar a una sola persona en dos circunstancias contrafácticas.) Este punto genera una advertencia. Es importante no pensar en los experimentos mentales como si estuvieran describiendo casos empíricos reales. Demos forma a esta observación.

Las clases de una teoría y sus principios de individuación, evolucionan como una respuesta al mundo tal como en realidad lo encontramos. Nuestras nociones de semejanza resultan de los intentos de explicar casos reales. No son necesariamente sensibles a ideales filosóficos preconcebidos.[10] Los términos de clase del discurso de las actitudes proposicionales son sensibles a semejanzas amplias y estables en el entorno real en que los agentes deben responder, operar y representar. Si la teoría hubiera sido confrontada con frecuencia, en entornos diferentes con agentes físicamente similares, podría haber desarrollado términos de clases diferentes. Sin embargo, estamos tan lejos de encontrar aproximaciones aun toscas de las semejanzas físicas globales entre los agentes,

10. Para una elaboración interesante de este tema en un contexto experimental, ver Amos Tversky, "Features of Similarity", *Psychological Review* 84 (1977), págs. 327-352. Cf. también Rosch y otros, *op. cit.*

que hay poca plausibilidad en la imposición de la similaridad física individual por sí misma, como una condición suficiente ideal para la identidad de los términos de clase en la psicología. Más aún, pienso que las semejanzas físicas locales entre las actividades psicológicamente relevantes de los agentes, están tan frecuentemente interconectadas con las constancias del entorno que una teoría psicológica que insistiera en abstraerse enteramente de la naturaleza del entorno al elegir sus términos de clases, estaría empíricamente arruinada.

El uso correcto de los contrafácticos en los experimentos mentales consiste en explorar el alcance y los límites de las nociones de clase que han sido desarrolladas con anterioridad al intentarse explicar casos empíricos reales. En el razonamiento contrafáctico suponemos una comprensión de lo que nuestro lenguaje expresa y exploramos sus condiciones de aplicación a través de la consideración de aplicaciones no reales. Los contrafácticos en los experimentos mentales filosóficos iluminan principios individuativos y teóricos con los cuales estamos ya comprometidos.

La implausibilidad empírica de los experimentos mentales es irrelevante con respecto a su intención filosófica, que concierne a la posibilidad, no a la plausibilidad. Los casos improbables aunque casos límite, son a veces necesarios para clarificar el *status* modal de las presuposiciones que gobiernan los ejemplos más mundanos. A la inversa, los casos altamente contrafácticos son, en gran medida, irrelevantes para la *evaluación* de una teoría empírica, excepto en los casos (no discutidos aquí) en los que presentan posibilidades empíricas que una teoría cuenta como imposibles. Invocar un principio filosófico general, como el principio de superveniencia, o insistir, dados los experimentos mentales, en que sólo ciertos tipos de similitud pueden ser relevantes para la psicología, sin criticar la teoría psicológica sobre bases empíricas o sin mostrar cómo las nociones de clase exhibidas por los experimentos mentales son empíricamente inadecuadas, es tratar las circunstancias contrafácticas como si fueran reales o bien caer en un *apriorismo* con respecto a la ciencia empírica.

Veamos ahora la forma más débil de la dificultad que hemos estado considerando. Ella pretende motivar un nuevo lenguaje individualista de atribución de actitudes. Como he advertido, aceptar tal lenguaje es consistente con el rechazo del (pan-) individualismo en la psicología. Hay en ella, una variedad de niveles o clases de explicación. Agregar otro, no alterará los problemas tratados aquí. Pero continuemos brevemente con el asunto.

En la psicología hay niveles de descripción individualista, por encima de lo fisiológico pero por debajo de lo actitudinal, que juegan un rol en las explicaciones sistemáticas. Se apela a procesos computacionales descriptos de manera formal en un intento de especificar un algoritmo mediante el cual la información proposicional de una persona es procesada. Pienso que para algunos propósitos, podría decirse que los protagonistas en nuestros experimentos mentales satisfacen algoritmos idénticos descriptos de manera formal. La información diferente se procesa de la "misma" manera, al menos en este nivel formal de descripción. Pero entonces, ¿no podríamos desear un nivel total de descripción entre el algoritmo formal y la adscripción ordinaria de actitudes proposicionales, que considere en todos los casos que la "información" es la misma entre los protagonistas en los experimentos mentales? Es una pregunta difícil y compleja que no intentaré responder aquí. Quiero, sin embargo, mencionar las razones para ser cautos acerca de una suplementación global de la psicología.

En primer lugar, la motivación para demandar las adiciones relevantes a la teoría psicológica, es empíricamente débil. En la literatura filosófica reciente, la motivación reside en gran parte, en intuiciones sobre los demonios cartesianos o en los cerebros en cubetas, cuya relevancia y aun coherencia ha sido repetidamente cuestionada; en preconceptos acerca de la superveniencia de lo mental en lo neural, que no tiene una garantía científica generalizada; en las aplicaciones erróneas de las observaciones ordinarias respecto de la causación; y en una concepción esquemática y poco clara de la conducta no basada en la práctica científica.[11] Por supuesto, uno puede investigar razonablemente cualquier hipótesis basado sólo intuitivamente en no más que una corazonada. Lo que es cuestionable es la opinión de que hay actualmente bases filosóficas y científicas fuertes para instituir un tipo nuevo de explicación individualista.

En segundo lugar, es fácil subestimar lo que está involucrado en la creación de un lenguaje individualista relevante que tenga un uso genui-

11. El más cuidadoso y plausible de varios artículos que defienden un lenguaje nuevo de la explicación individualista, es el de Stephen White, "Partial Character and the Language of Thought", *Pacific Philosophical Quarterly* 63 (1982), págs. 347-365. Me parece, sin embargo, que la mayoría de los problemas mencionados en el texto, aquí y más abajo, bloquean esa defensa. Más aún, las tareas positivas establecidas para el nuevo lenguaje han sido ya desarrolladas por el lenguaje actual no-individualista en la psicología. Las intuiciones relativas a los cerebros en cubeta son puntos muy complejos que no puedo tratar aquí. Los discuto en "Cartesian Error and the Objectivity of Perception", *op. cit.*

no en la psicología. Las explicaciones de tal lenguaje han sido, hasta aquí, muy provisionales. No alcanza con esquematizar una semántica que diga, en efecto, que una oración resulta verdadera en todos los mundos en los que los protagonistas, químicamente idénticos, en experimentos mentales relevantes, no puedan distinguirla. Tal explicación no brinda reglas claras para el *uso* del lenguaje, y mucho menos brinda una demostración de que pueda cumplir una función distintiva en la psicología. Más aún, la explicación del lenguaje (o del componente del lenguaje) individualista, sólo para el caso especial en el cual los estados funcionales (especificados de forma individualista) o fisiológicos del usuario del lenguaje sean constantes, es psicológicamente inútil, puesto que dos personas nunca son realmente idénticas en sus estados físicos.

Para idear un lenguaje individualista no sirve limitar su referencia a las propiedades objetivas accesibles a la percepción. Porque nuestro lenguaje para adquirir nociones de propiedades físicas perceptualmente accesibles no es individualista. En términos más generales, como he argumentado en otro lado (ver la última *op. cit.*, nota 1), cualquier actitud que contenga nociones para propiedades, eventos y objetos físicos es no individualista.[12] Las suposiciones respecto de la representación objetiva, que se necesitan para generar el argumento, son bastante mínimas. Pienso que es cuestionable que haya una concepción coherente de la representación objetiva que pueda apoyar un lenguaje individualista de la atribución de actitud intencional. Quienes abogan por tal lenguaje deben explicar tal concepción en profundidad o bien atribuir estados intencionales que carezcan de una referencia física objetiva.

Traductores: Eduardo Barrio, Patricia Brunsteins, Margarita Roulet y Julia Vergara.
Revisión técnica: Eduardo Rabossi.

12. Ver especialmente "Intellectual Norms and Foundations of Mind", *op. cit.*, pero también "Individualism and the Mental", *op. cit.*, págs. 81-82.

B. Mundos nocionales, semántica conceptual e individualismo

Capítulo 10

MÁS ALLÁ DE LA CREENCIA *
(SELECCIÓN)

Daniel Dennett

[...]

IV. Actitudes Nocionales

Dadas las objeciones de Putnam y de otros [autores] a las actitudes proposicionales "clásicas", hemos reparado en la siguiente pregunta: ¿cuál es la contribución del organismo [*organismic*] en la fijación de las actitudes proposicionales? La respuesta caracterizaría los estados psicológicos "en el sentido estrecho" ["*narrow sense*"]. El intento de caracterizar esos tipos de estados psicológicos restringidos como *actitudes oracionales* [*sentential attitudes*] se topó con varios problemas, el más importante de los cuales fue que cualquier caracterización de la actitud oracional, al ser esencialmente un tipo de molde [*casting*] *sintáctico*, producía cortes demasiado finos. Respecto del experimento mental de Putnam concedimos que la duplicación *física* es suficiente pero no necesaria para la identidad de la contribución del organismo; también podríamos garantizar que la similaridad [*similarity*] más débil capturada por la duplicación *sintáctica* (en algún nivel de abstracción) sería suficiente para la identidad de la contribución del organismo, pero aun cuando la identidad de la contribución del organismo —la "gemelidad" psicológica-estrecha [*narrow-psychological twinhood*]— resulte ser una condición muy severa, no parecería requerir la gemelidad sintáctica, en ningún nivel de la descripción. Considérese la pregunta, en algún sentido análoga: ¿comparten una descripción sintáctica (esto es, una tabla de máquina) todas las máquinas de Turing que computan una misma función? No, a menos que ajustemos nuestros niveles de descripción de la tabla de máquina y de la conducta *input-output* de modo tal que se

* "Beyond Belief", en Andrew Woodfield (comp.), *Thought and Object*, Oxford, Clarendon Press, 1982, págs. 36-60. Con autorización del autor y de Clarendon Press.

asocien de manera trivial. Qué debería *contar* como equivalencia para las máquinas de Turing (o para los programas de computación) es una cuestión discutida; no lo sería si no fuera por el hecho de que las descripciones no-trivialmente diferentes expresadas en términos de la "sintaxis" interna, pueden producir la *misma* "contribución", en algún nivel de descripción útil.

La analogía es imperfecta, sin duda, y otras consideraciones —por ejemplo, consideraciones biológicas— podrían pesar en favor de la suposición de que la gemelidad psicológica-estrecha *completa* ha requerido la gemelidad sintáctica en algún nivel, pero aun si se garantizara esto, no se seguiría que la similaridad psicológica *parcial* pudiera ser siempre descripta en algún sistema general de descripción sintáctica aplicable a todos aquellos que comparten el rasgo psicológico. La gente que es vanidosa, o paranoica, por ejemplo, seguramente es similar desde un punto de vista psicológico; en cada caso, una gran parte de la similaridad parecería bien captada al hablar de creencias similares o compartidas. Aun si uno adopta una línea rigurosa autodestructiva [*self-defeatingly*] de la identidad de las creencias (de acuerdo con la cual nunca un par de personas comparte realmente una creencia), estas *similaridades* en las creencias exigen ser captadas por la psicología. No podría sostenerse plausiblemente que dependen del monolingüismo [*monolingualism*]; [afirmando que] los cerebros de las personas vanidosas hablan el mismo mentalés [*Mentalese*]. Tampoco podemos captar esas similaridades en el estado-de-creencia [*belief-state*] *via* actitudes *proposicionales*, a causa del carácter deíctico [*indexicality*] de muchas de las creencias cruciales: "La gente *me* admira", "La gente trata de arruinar*me*".

Estas consideraciones sugieren que lo que estamos tratando de caracterizar es una posición intermedia; podría decirse, a mitad de camino entre la sintaxis y la semántica. Llamémosla *psicología de la actitud nocional* [*notional attitude psychology*]. Queremos que haga [*work out*] que yo y mi *Doppelgänger* —y cualquier otro par de gemelos psicológico-restringidos— tengamos exactamente las mismas actitudes nocionales, de tal manera que nuestras diferencias en las actitudes proposicionales se deban enteramente a las diferentes contribuciones ambientales. Pero también queremos que haga que usted y yo, que no somos gemelos psicológicos sino "de un mismo parecer" ["*of like mind*"] en muchos temas, compartamos una variedad de actitudes nocionales.

Una idea familiar que ha aparecido bajo diferentes disfraces puede

ser adaptada para nuestros propósitos: la idea del mundo subjetivo de una persona; *The World I Live In*, de Helen Keller, o *The World According to Garp*, de John Irving, por ejemplo. Tratemos de caracterizar el *mundo nocional* de un sujeto psicológico, de manera tal que, por ejemplo, aunque mi *Doppelgänger* y yo vivamos en mundos reales diferentes —la Tierra Gemela y la Tierra— tengamos el *mismo* mundo nocional. Usted y yo vivimos en el mismo mundo real, pero tenemos mundos nocionales diferentes, aunque haya una considerable superposición entre ellos.

Un mundo nocional debería verse como una suerte de mundo de *ficción* [*fictional world*] inventado por un teórico, un observador externo [*third-party*], para caracterizar los estados psicológico-estrechos de un sujeto. Puede suponerse que un mundo nocional está repleto de objetos nocionales, y que es el escenario de eventos nocionales; podría decirse, de todos los objetos y eventos en los que el sujeto cree. Si suavizamos nuestro solipsismo metodológico por un momento, nos daremos cuenta de que algunos objetos del mundo real habitado por el sujeto "concuerdan" ["*match*"] con objetos en el mundo nocional del sujeto, pero que otros no. El mundo real contiene muchas cosas y eventos que no tienen contraparte en el mundo nocional de ningún sujeto (excluyendo el mundo nocional de un Dios omnisciente), y los mundos nocionales del crédulo o de los sujetos confundidos u ontológicamente licenciosos, contendrán objetos nocionales que no tienen contrapartes en el mundo real. La tarea de describir las relaciones que pueden existir entre las cosas en el mundo real y las cosas en el mundo nocional de alguien, está, notoriamente, plagada de enigmas [*puzzle-ridden*]; ésta es una razón para cobijarse en el solipsismo metodológico: para dejar a un lado [*factor out*] temporariamente estos temas problemáticos.

Nuestra decisión nos ha hecho aterrizar en un territorio muy familiar: ¿qué son los objetos nocionales sino los *objetos intencionales* de Brentano? El solipsismo metodológico es, aparentemente, una versión de la *epojé* de Husserl, el poner entre paréntesis. ¿Puede ser que la alternativa tanto a la psicología de la actitud proposicional como a la psicología de la actitud oracional sea... la Fenomenología? No precisamente. Hay una gran diferencia entre el enfoque esbozado aquí y los enfoques tradicionales asociados a la Fenomenología. Mientras que los fenomenólogos sostienen que uno puede acceder a su *propio* mundo nocional a través de alguna muestra introspeccionista de gimnasia mental —llamada, por algunos, la reducción fenomenológica— a nosotros nos interesa determinar el mundo nocional de *otro*, desde fuera. La tra-

dición de Brentano y de Husserl es la *autofenomenología*; yo estoy proponiendo la *heterofenomenología*.[16] Aunque los resultados puedan tener un parecido sorprendente, las suposiciones que los hacen posibles son muy diferentes.

La diferencia puede verse mejor con la ayuda de la distinción, recientemente resucitada por Fodor (1980), entre lo que llama, siguiendo a James, psicología *naturalista* [*naturalistic*] y *racional*. Fodor cita a James:

> En conjunto, pocas fórmulas recientes han prestado más servicio de fondo [*of a rough sort*] en psicología que la fórmula spenceriana de que la esencia de la vida mental y de la vida corporal son una, a saber, "el ajuste de las relaciones internas con las externas". Tal fórmula es la encarnación de la vaguedad, pero debido a que toma en cuenta el hecho de que las mentes habitan medios que actúan sobre ellas y ante los cuales ellas reaccionan a su vez; debido a que, en pocas palabras, toma a la mente en medio de todas sus relaciones concretas, es inmensamente más fértil que la "psicología racional", pasada de moda, que trataba al alma como un[a entidad] existente separada, suficiente por sí misma, de la que se proponía considerar sólo su naturaleza y sus propiedades (James, 1890, pág. 6).

James canta loas a la psicología naturalista, la psicología en el sentido *amplio*, pero la moraleja de la Tierra Gemela, extraída explícitamente por Fodor, es que la psicología naturalista tiene pretensiones demasiado amplias para ser factible. Los fenomenólogos extraen la misma conclusión, aparentemente, y ambas se tornan versiones diferentes del solipsismo metodológico: se preocupan por el sujeto psicológico "como un[a entidad] existente separada, suficiente por sí misma", pero cuando "consideran su naturaleza y sus propiedades", ¿qué encuentran? Los fenomenólogos, usando algún tipo de introspección, pretenden encontrar *lo dado* en la experiencia, que resulta ser la materia prima [*raw material*] para su construcción de mundos nocionales. Si Fodor, usando algún tipo de inspección interna (imaginada) de la maquinaria, alegó encontrar un texto en mentalés *dado* en el *hardware* (que se constituiría en la materia prima para la construcción de las actitudes nocionales del sujeto), tendríamos tanta razón para dudar de la existencia de lo dado en este caso como en el caso de la Fenomenología. James está en lo correcto: uno no puede hacer *psicología* (como opuesto a, diga-

16. Véase mi "Two Approaches to Mental Images" (Dennett, 1978), para algunas vicisitudes de la auto-fenomenología. Véase también Campbell (1977).

mos, neurofisiología) sin determinar las propiedades *semánticas* de los eventos y de las estructuras internas bajo examen, y uno no puede descubrir las propiedades semánticas sin prestar atención a las relaciones de esos eventos o estructuras internas con cosas en el entorno del sujeto. Pero en ningún lado está escrito que el entorno relativo al cual fijamos las propiedades semánticas de un sistema, deba ser el entorno *real* o el entorno *efectivo* [*actual*] en el cual el sistema ha crecido. Un entorno de ficción, un entorno idealizado o imaginario podría funcionar igual de bien. La idea es que para hacer una teoría de la "representación mental", es necesario hacer la semántica de las representaciones *desde el principio*. (Uno no puede hacer primero la sintaxis y después la semántica.) Pero eso significa que se necesita un modelo, en el sentido de la semántica tarskiana. Un modelo de ficción, sin embargo, podría permitir una semántica tarskiana suficiente como para determinar la semántica parcial, o proto-semántica, que necesitamos para caracterizar la contribución del organismo.

La idea de un mundo nocional, entonces, es la idea de un modelo —pero no necesariamente el modelo efectivo, real, verdadero— de las representaciones internas de uno. *No consiste él mismo en representaciones sino en representados* [*representeds*]. Es el mundo "en el que vivo", no el mundo de las representaciones *en mí*. (Hasta aquí, esto es Brentano puro, al menos como yo lo entiendo. Véase Aquila, 1977.) El teórico que desea caracterizar los estados psicológico-restringidos de una criatura, o con otras palabras, la contribución del organismo de esa criatura a sus actitudes proposicionales, *describe* un mundo de ficción; la descripción existe en el papel, el mundo de ficción no existe, pero los habitantes del mundo de ficción son tratados como los referentes nocionales de las representaciones del sujeto, como los objetos intencionales de ese sujeto. Se espera que por medio de esa trama el teórico pueda obtener los beneficios del naturalismo de James y Spencer sin las dificultades planteadas por Putnam y los otros.

Entonces, la pregunta es: ¿qué guía nuestra construcción del mundo nocional de un organismo? Supongamos, para dramatizar el problema, que recibimos una caja que contiene un organismo, de no sabemos dónde, vivo pero congelado (o comatoso), y por lo tanto separado de su entorno. Tenemos una fotografía instantánea laplaceana del organismo —una descripción completa de su estructura y de su composición interna— y podemos suponer que esto nos permite determinar exactamente cómo *respondería* a los nuevos impactos del entorno, si lo libráramos de su estado de animación suspendida y de su aislamiento.

Nuestra tarea es como el problema planteado cuando se nos muestra un artefacto extraño o antiguo y se nos pregunta: ¿para qué sirve? ¿Es una máquina de hacer agujas, o un artefacto para medir la altura de los objetos distantes, o un arma? ¿Qué podemos aprender del estudio del objeto? Podemos determinar cómo se combinan sus partes, qué sucede bajo varias condiciones, y demás. Podemos también buscar melladuras e incisiones delatoras, el desgaste natural. Una vez que hemos compilado estos hechos, tratamos de imaginar una situación en la cual, dados esos hechos, [el artefacto] realizaría *de manera óptima* alguna función útil imaginable. Si el objeto fuera igualmente bueno para coser velas [*sails*] o para deshuesar cerezas, no seríamos capaces de decir lo que *realmente* es —*para* qué es— sin saber de dónde vino, quién lo hizo, y por qué fue hecho. Todos estos hechos podrían haberse desvanecido sin dejar rastros. La identidad verdadera del objeto, su esencia, podría resultarnos totalmente indeterminable, no importa cuán asiduamente lo hayamos estudiado. Esto no significa que no haya una cuestión de hecho [*fact of the matter*] acerca de si la cosa era para deshuesar cerezas o para coser velas, sino que la verdad, sea la que fuere, ya no es relevante. Sería uno de esos hechos históricos ociosos o inertes, como el hecho de que algo del oro que está en mis dientes perteneció o no una vez a Julio César.

Enfrentados con nuestro nuevo organismo, podemos con bastante facilidad determinar para qué es —es para sobrevivir y florecer y reproducir su especie— y no tendríamos mayor problema en identificar sus órganos sensoriales y sus modos de acción y sus necesidades biológicas. Dado que, *ex hypothesi*, podemos calcular qué haría si... (para todos los antecedentes que completemos), podemos determinar, por ejemplo, que comerá manzanas pero no pescado, que tratará de evitar los lugares muy iluminados, que está dispuesto a hacer ciertos ruidos bajo ciertas condiciones, etcétera. Ahora bien, ¿qué tipo de entorno encajaría con estos talentos y propensiones? Cuanto más sabemos acerca de la estructura interna, de las disposiciones conductuales y de las necesidades sistémicas del organismo, más particular se vuelve nuestro entorno hipotético ideal. Por "entorno ideal" no me refiero al mejor de todos los mundos posibles para ese organismo ("...los pajarillos cantan, la luna se levanta...") sino al entorno (o clase de entornos) en los cuales el organismo, como está constituido actualmente, encaja mejor. Podría ser un mundo francamente terrible, pero al menos el organismo está preparado para habérselas con él. Podemos aprender algo acerca de los enemigos del organismo —reales o sólo nocionales— notando su colo-

ración protectora o su conducta de escape o ... cómo respondería a ciertas cuestiones.

En la medida en que el organismo con el que estamos tratando es muy simple y tiene, por ejemplo, poca o ninguna plasticidad en su sistema nervioso (de modo que no puede aprender), el límite de la especificidad del entorno ideal imaginado puede no llegar a distinguir entornos radicalmente diferentes pero que encajan igualmente bien, como en el caso del artefacto. En la medida en que la capacidad de aprender y recordar crece, y en la medida en que la riqueza y la complejidad de las relaciones posibles con las condiciones del entorno crecen, [17] la clase de modelos igualmente aceptables (entornos ideales hipotéticos) disminuye. Más aún, en criaturas con la capacidad de aprender y de almacenar en la memoria información sobre su mundo, un principio exegético nuevo y más poderoso entra en juego. Las melladuras e incisiones en el deshuesador de cerezas (¿o era para coser velas?) pueden en alguna ocasión probar ser delatoras, pero las melladuras e incisiones en la memoria de la criatura que aprende, están diseñadas para ser delatoras, para grabar con alta fidelidad tanto los encuentros particulares como las lecciones generales, para el uso futuro. Dado que las melladuras e incisiones de la memoria son para el uso futuro, podemos esperar "leerlas" explotando nuestro conocimiento de las disposiciones que dependen de ellas, en la medida en que supongamos que las disposiciones así asignadas, son, en general, apropiadas. Tales interpretaciones de los "rastros de memoria" dan información más específica sobre el mundo en el cual la criatura vivió y al cual se ha acomodado. Pero no seremos capaces de dar información sobre este mundo a partir de la desinformación [*misinformation*], y así el mundo que extrapolamos como *constituido* por el estado actual del organismo será un mundo ideal, no en el sentido de *mejor*, sino en el sentido de *no real* [*unreal*].

Los naturalistas insistirán, correctamente, en que el entorno efectivo, tal como se encontró, ha dejado su marca sobre el organismo y lo ha conformado intrincadamente; el organismo está en su estado actual *por la historia que ha tenido*, y sólo una historia tal podría de hecho haberlo puesto en su estado presente. Pero a modo de experimento mental, podemos imaginar que creamos un duplicado cuya historia *aparente* no es su historia efectiva (como en el caso de una antigüedad falsa, con sus marcas y desgaste natural simulados). Una duplicación completa tal (la cual es sólo lógicamente posible en un experimento mental) es el

17. Véase "True believers" (Dennett, 1987).

caso límite de algo efectivo y familiar: cualquier rasgo particular de un estado actual puede ser ilegítimo [*misbegotten*], de forma tal que la manera en que el mundo debería haber sido para la criatura que ahora está en este estado mental, no es exactamente la manera en que el mundo fue. El mundo nocional que describimos por extrapolación a partir del estado actual no es exactamente el mundo que consideramos que ha creado ese estado, aun cuando conocemos ese mundo efectivo; es más bien el mundo aparente [*apparent*] de la criatura, el mundo aparente para la criatura, tal como se manifestó en el estado disposicional total actual de la criatura.

Supongamos que aplicamos este ejercicio de imaginación en la formación de un mundo nocional para organismos altamente adaptativos [*adaptive*] como nosotros mismos. Tales organismos tienen estructura interna y rasgos disposicionales tan ricos en información acerca del entorno en el cual se desarrollaron que podríamos en principio decir: este organismo se ajusta mejor al entorno en el cual hay una ciudad llamada Boston, y en el cual el organismo ha pasado su juventud en la compañía de organismos llamados... y así en más. No seríamos capaces de distinguir a Boston de Boston-en-la-Tierra-Gemela, por supuesto, pero excepto por tales variaciones virtualmente indistinguibles sobre el tema, nuestro ejercicio en la formación de un mundo nocional terminaría en una única solución.

Éste es, de cualquier manera, el mito. Es un mito prácticamente inútil, por supuesto, pero teóricamente importante, porque revela las suposiciones fundamentales que se hacen acerca de la dependencia última de la contribución del organismo en su constitución física. (Está dependencia es conocida de otra manera como la superveniencia [*supervenience*] de los rasgos psicológicos (restringidos) sobre los rasgos físicos; ver, por ejemplo, Stich, 1978.) Al mismo tiempo, el mito preserva la *sub*determinación [*underdetermination*] de la referencia última, que fue la moraleja extraída de las consideraciones de Putnam. Si hay un lenguaje del pensamiento, es así como uno tendría que ajustar el modo de descubrirlo y traducirlo, sin los beneficios de los intérpretes bilingües o de la evidencia circunstancial acerca del origen del texto. Si hay alguna alternativa de tercera persona para el método del fenomenólogo, dudosamente introspeccionista (*genuinamente* solipsista), si la hetero-fenomenología es posible en absoluto, tendremos que seguir ese método.

En principio, entonces, los últimos frutos del método, aplicado a un ser humano bajo las restricciones del solipsismo metodológico, serían ciertamente una descripción exhaustiva del mundo nocional de una per-

sona, completo con sus identidades erróneas, quimeras y espectros, errores fácticos y distorsiones.[18] Podemos pensarlo como *el* mundo nocional de un individuo, pero por supuesto [aun] la mayor descripción exhaustiva fallaría al especificar un único mundo. Por ejemplo, las variaciones en un mundo [que esté] enteramente más allá del alcance de la vista o de los intereses de una persona, generarían diferentes mundos posibles igualmente consistentes con la determinación maximal provista por la constitución de la persona.

La situación es análoga a la de los más familiares mundos de ficción, tales como el mundo de Sherlock Holmes o el Londres de Dickens. Lewis (1978) provee una explicación de "verdadero en ficción", la semántica de la interpretación de la ficción, que desarrolla la idea que necesitamos: "el" mundo de Sherlock Holmes es mejor concebido, formalmente, como un *conjunto* de mundos posibles, aproximadamente: todos los mundos posibles consistentes con el *corpus* entero de textos de Sherlock Holmes [escritos] por Conan Doyle.[19] Igualmente, "el"

18. ¿Y qué decir de los objetos de sus temores, ilusiones y deseos? ¿Son ellos habitantes de los mundos nocionales del sujeto, o debemos añadir un mundo de deseos, un mundo de miedos, y así sucesivamente, al mundo creencial del sujeto? (Joe Camp y otros han insistido en este problema). Cuando algo que el sujeto cree que existe es también temido o deseado por él, no hay problema: algún habitante de su mundo nocional es simplemente teñido [*coloured*] con deseos, miedos o admiración, o lo que sea. Cómo tratar "La casa soñada que algún día deseo construir" es otro problema. Posponiendo los detalles para alguna otra ocasión, aventuraría, sin cautela, alguna afirmación general. Mi casa soñada no es un habitante de mi mundo nocional a la par con mi casa o incluso con la casa que terminaré en vida; pensar acerca de ello (mi casa soñada) no tiene que ser, por ejemplo, analizado de la misma manera que pensar acerca de mi casa o de la casa que terminaré en vida... Mi casa soñada logra constituirse indirectamente en mi mundo nocional *vía* lo que podríamos llamar mis *especificaciones* [*specifications*], que son habitantes perfectamente ordinarios de mi mundo nocional, y mis creencias generales y otras actitudes. Creo en mis especificaciones, que ya existen en el mundo como ítemes de piezas mentales creadas por mi pensamiento, y hay entonces creencias y deseos generales, etcétera, que involucran esas especificaciones: decir que mi casa soñada es construida de cedro no es decir que mis especificaciones están hechas de cedro, sino decir que cualquier casa construida por mis especificaciones sería hecha de cedro. Decir que deseo construir*la* el año próximo es decir que planeo construir una casa de acuerdo con mis especulaciones [*to my specs*] el año próximo.

19. Las características especiales de la ficción (literaria) llevan a Lewis a hacer modificaciones sustanciales e ingeniosas a esta idea, para dar cuenta del rol de las suposiciones subyacentes, el conocimiento del narrador y demás, en la interpretación normal de ficción. Por ejemplo, suponemos que el mapa de Holmes de la ciudad de Londres es el de la Londres victoriana, excepto en aquello prescindido por las invenciones de Conan Doyle; los textos no afirman ni implican estrictamente que Holmes no tenga una tercera ventana de la nariz, pero los mundos posibles en los cuales esto es el caso, son excluidos.

mundo nocional que describimos podría ser mejor visto, formalmente, como el conjunto de todos los mundos posibles consistentes con la descripción maximal.[20] Adviértase que la descripción es la descripción del *teórico*; no *supongamos* que las características estructurales del organismo sobre las que el teórico basa su descripción incluyen elementos que son ellos mismos descripciones. (Las características del deshuesador de cerezas que nos lleva a describir una cereza [más que un durazno o una aceituna] no son ellas mismas descripciones de las cerezas.) Desde esta perspectiva, podemos ver que Putnam ha concebido a la Tierra y a la Tierra Gemela como miembros del conjunto de los mundos posibles que *es* el mundo nocional que comparto con mi *Doppelgänger*. XYZ aplaca la sed, disuelve el engrudo del empapelado y produce el arco iris tan bien como lo hace H_2O; su diferencia con H_2O está por debajo de todos los umbrales de discriminación míos y de mi *Doppelgänger*, a condición de que, presumiblemente, ninguno de nosotros sea, o consulte con, un químico astuto o un microfísico.

Dado un mundo nocional para un sujeto, podemos hablar respecto de aquello *acerca* de lo que son las creencias del sujeto, en un sentido peculiar, pero familiar de "acerca de". Goodman (1961) discute oraciones de Dickens que son "acerca de Pickwick"; una característica semántica de estas oraciones es que no son genuinamente relacionales, porque no hay un Mr. Pickwick acerca del cual ellas son, en un sentido relacional fuerte. Con un espíritu similar, Brentano discute el *status* "relacional" ["*relationlike*"] de los fenómenos mentales cuyos objetos intencionales son no-existentes (véase Aquila, 1977). Una suposición apta de la psicología de la actitud nocional es que el teórico puede usar ser-acerca-de-Pickwick [*Pickwick-aboutness*] y sus "parientes", como las propiedades semánticas que uno necesita para los fundamentos de cualquier teoría de la representación mental.

La estrategia no es nueva. Aunque la psicología de la actitud nocional ha sido diseñada aquí como una respuesta a los problemas filosóficos con que tropiezan las actitudes proposicionales y la psicología de la actitud oracional, puede fácilmente discernirse como la metodología tácita y la ideología de la principal rama de la Inteligencia Artificial.

20. Se ha prestado mucha atención a distintas variaciones de la idea de tratar un "mundo-creencial" ["*belief-world*"] como el conjunto de todos los mundos posibles en los que las creencias de uno son verdaderas, desde que la idea fue introducida por Hintikka (1962). Ordenar esas variaciones, compararlas y contrastarlas con mi desarrollo del tema es un trabajo que dejo para otra ocasión.

[...]

La elaboración del entorno imaginario ideal para propósitos de comparación de los sistemas internamente diferentes es una estrategia difundida, por ejemplo, en ingeniería. Podemos comparar el poder de los motores de automóviles diferentes imaginando que tiran en una contienda con un cierto caballo de ficción, o podemos comparar su eficiencia en el consumo de combustible viendo hasta dónde empujarán un auto en un cierto entorno simulado. El uso de un entorno ideal permite describir similitudes o capacidades *funcionales* independientemente de los detalles de implementación o de actuación. Utilizar la estrategia en la psicología para elaborar los mundos nocionales es un caso particularmente complejo. Nos permite describir similitudes parciales en las "capacidades" psicológicas de diferentes sujetos —por ejemplo, sus *poderes* representacionales— en forma tal que somos neutrales acerca de su implementación; por ejemplo, de sus *medios* [*means*] representacionales.

La analogía con la ficción es de nuevo útil para señalar este punto. ¿Cuál es exactamente la similitud entre *Romeo y Julieta*, de Shakespeare, y *West Side Story* de Bernstein? Sabemos que la última estuvo "basada en" la primera, pero ¿qué tienen de hecho en común? ¿Son acerca de las mismas personas? No, ya que ambas son ficciones. ¿Contienen las mismas o similares representaciones? ¿Qué podría significar esto? ¿Tienen las mismas palabras, oraciones o descripciones? Los argumentos de ambas están escritos en inglés, pero esto es claramente irrelevante, porque la similaridad que buscamos sobrevive después de la traducción a otros lenguajes y es evidente —de manera dramática— en el film de *West Side Story* y en la ópera de Gounod. La similaridad es independiente de cualquier *medio* particular de representación —el guión, los diálogos, las descripciones, los actores en escena o frente a las cámaras— y concierne a *lo que es representado*. No es cualquier tipo de similaridad *sintáctica*. Ya que tales similaridades son tan evidentes en la ficción como en los informes fácticos, debemos comprender "lo que se representa" recurriendo a los elementos de un mundo nocional y no necesariamente del mundo real. Podemos comparar diferentes mundos nocionales o ficcionales con respecto a asuntos importantes o menudos, tal como podemos comparar diferentes partes del mundo real. Podemos comparar un mundo nocional con el mundo real. (El mundo nocional del miope Mr. Magoo sólo se asemeja al mundo real de manera intermitente y parcial, pero sólo lo suficiente —milagrosamente— como para salvarlo del desastre.)

¿Cuándo diremos, entonces, que dos personas diferentes comparten una actitud nocional o un conjunto de actitudes nocionales? Cuando sus mundos nocionales tengan un punto o una región de similaridad. Los mundos nocionales son centrados-en-el-agente [*agent-centered*], son egocéntricos (Perry, 1977; Lewis, 1979); cuando se comparan los mundos nocionales buscando la similitud psicológica será útil, típicamente, "superponer" por lo tanto los centros [*centers*] —para que los orígenes, la intersección de los ejes, coincidan—, antes de testear la similaridad. De esta manera las similitudes psicológicas de dos personas paranoicas emergerán, mientras la diferencia psicológica entre el masoquista y su pareja sádica se mantendrá a pesar de la gran similaridad en las *dramatis personae* de sus mundos nocionales cuando se las ve descentradas [*uncentered*].[21]

Las perspectivas de un método riguroso de comparación de mundos nocionales —un procedimiento de decisión para encontrar y valorar, por ejemplo, puntos de coincidencia— es débil. Pero siempre hemos sabido que las perspectivas para establecer condiciones de identidad de actitudes proposicionales son igualmente débiles. Creo que la sal es cloruro de sodio, pero mi conocimiento de química es pobre; el químico también cree que la sal es cloruro de sodio, pero no hay ninguna manera vigorosa [*crisp*] de captar el núcleo común de nuestras creencias. (Dennett, 1969). La comparabilidad de creencias, vistas como actitudes nocionales o como actitudes proposicionales, no va a volverse rutinaria por un golpe de suerte teórico. La *ganancia en precisión*, que uno podría equivocadamente haber esperado obtener aislando y traduciendo "el lenguaje del pensamiento" —si es que existe—, no mejorará la comparabilidad de las *creencias*, tales como la del químico y la mía acerca de la sal, sino sólo la comparabilidad de un cierto tipo novedoso de oraciones, las oraciones que están en la cabeza. Pero las oraciones siempre fueron sutilmente comparables. El químico que habla español y yo usamos exactamente las mismas palabras para expresar nuestras creencias acerca de la sal, y si por casualidad nuestros cerebros también lo hacen, tenemos aun el mismo problema de comparabilidad de nuestras creencias.

Un lenguaje del pensamiento no daría más ventajas [*leverage*] en el

21. Las cuestiones relativas a "yo" y a la deíxis son mucho más complicadas que lo que revela este apresurado reconocimiento. Véase no sólo Perry y Lewis, sino también Castañeda (1966, 1967, 1968). Para iluminar las reflexiones sobre temas similares, véase Hofstadter, 1979, págs. 373-76.

controvertido caso de las creencias irracionales —y especialmente de las contradictorias—, y por la misma razón. Supongamos que se expone la hipótesis diciendo que Bill tiene un par particular de creencias contradictorias: cree al mismo tiempo que Tom es confiable y que Tom no es confiable. En cualquier lenguaje que valga la pena lo que se requiere es determinar cuándo una oración contradice a la otra; por eso, conociendo el lenguaje del pensamiento de Bill, buscamos en su cerebro el par relevante de oraciones. ¡Y las encontramos! ¿Qué mostraría esto? La pregunta aún seguiría en pie: ¿cuál [oración] él cree (si cree alguna)? Podríamos encontrar, investigando más, que una de esas oraciones era un vestigio y no era funcional; nunca fue borrada de la pizarra cerebral, pero nunca fue consultada, tampoco. O podríamos encontrar que una oración (la oración en el mentalés que corresponde a "Tom no es de confiar") fue consultada con intermitencia (y llevó a actuar), una buena evidencia de que Bill cree que Tom no es de confiar, pero que pasa inadvertida en su mente. Él olvida, y entonces su *bonhomía* natural se hace cargo y creyendo que la gente en general es de confiar, se comporta como si creyera que Tom es de confiar. O quizás encontremos una conducta verdaderamente conflictiva en Bill: habla y habla acerca de la confiabilidad de Tom, pero advertimos que nunca le vuelve la espalda. Se pueden multiplicar los casos, llenando huecos y ampliando los extremos, pero en *ninguno* de los casos la presencia o ausencia de contradicción explícita en el mentalés juega más que un rol de soporte periférico en nuestra decisión de caracterizar a Bill como vacilante, olvidadizo, indeciso, o verdaderamente irracional. La conducta de Bill cuenta mucho, pero la conducta no *resuelve* el problema tampoco.

La gente desde luego se confunde y aún peor, algunas veces se vuelve bastante loca. Decir que alguien es irracional es decir (en parte) que en algún respecto está mal equipado para tener trato con el mundo que habita; que encaja mal en su nicho. En los peores casos podemos ser incapaces de inventar algún mundo nocional para él; ningún mundo posible sería un lugar en donde él encaje bien. Se podría dejar el asunto aquí, o se podría tratar de ser más descriptivo respecto de la confusión de la persona.[22] Se podría componer una descripción declaradamente

22. "Un hombre puede pensar que cree *p*, mientras que su conducta sólo puede ser explicada por la hipótesis de que cree que *no-p*, dado que se sabe que desea *z*. Quizá la confusión en su mente no pueda ser expuesta por ninguna explicación simple [o compleja —D.C.D.] de lo que cree: quizá sólo una reproducción de la complejidad y confusión será correcta" (Hampshire, 1975, pág. 123).

inconsistente, citando minuciosamente las propensiones conductuales del sujeto y su constitución interna, en apoyo de las distintas partes de la descripción. No podría decirse que esa descripción inconsistente sea *un* mundo nocional, ya que los mundos nocionales, como conjuntos de mundos *posibles*, no pueden tener propiedades contradictorias, pero nada garantiza que el sujeto tenga un único mundo nocional coherente. Su mundo nocional puede ser transformado en mundos fragmentarios, solapados [*overlapping*], en competencia.[23] Cuando el teórico, el heterofenomenólogo o el psicólogo de los mundos nocionales toma la decisión de ofrecer una descripción expresamente inconsistente de un mundo nocional, ella no cuenta como una caracterización establecida, positiva, de un mundo nocional, sino como un rendirse ante la confusión, abandonando el intento de una interpretación (completa). Es análogo al error en la cita directa, cuando se expresan los comentarios de alguien. "Bien, lo que él *dijo* fue: 'La nada nadea' ".

La heterofenomenología de los mundos nocionales, entonces, no resuelve las discusiones e indeterminaciones, y aun agudiza los límites de los pensamientos cotidianos [*everyday-folks' thinking*] acerca de la creencia; ella hereda los problemas y sólo los reformula en un formato ligeramente nuevo. Se podría muy bien preguntar qué recursos la hacen recomendable. La perspectiva de construir el mundo nocional de una criatura efectiva a partir del examen de su constitución física es tan remota como pueda pensarse, así que, ¿qué valor puede haber en la concepción del mundo nocional de una criatura? Trabajando en otra dirección: comencemos con la descripción de un mundo nocional y entonces preguntemos cómo diseñar una "criatura" con ese mundo nocional. Parte del talante de la Inteligencia Artificial es proveer una manera de comenzar con las que son esencialmente categorías y distinciones esencialmente fenomenológicas —características de los mundos nocionales— y trabajar hacia atrás en busca de hipótesis acerca de cómo implementar esas competencias.[24] Se comienza con *poderes* [*powers*] representacionales y se trabaja hacia los *medios* [*means*]. Los filósofos han jugado también con esta estrategia.

La literatura filosófica reciente sobre la distinción entre las creen-

23. Véase la discusión de la Fenomenología y de la "Feenomanología" ["*Feenomanology*"] en "Two Approaches to Mental Images", Dennett (1978). Véase también los comentarios de Lewis (1978) acerca de cómo tratar la inconsistencia en un trabajo de ficción.

24. Putnam (1975) describe la teoría del significado apta para el solipsismo metodológico como una teoría de la "competencia individual" (pág. 246).

cias *de re* y *de dicto*, y otras actitudes, está repleta de sugerencias incompletas para varios géneros de la maquinaria mental que podría jugar un rol crucial en bosquejar esa distinción: los *nombres vívidos* [*vivid names*] de Kaplan (1968), los *modos de presentación* de Schiffer (1978) y de varios otros autores, y los *aspectos* de Searle (1979), para nombrar algunos. Se supone que éstos son típicamente definibles sólo en términos de *psicología estrecha*;[25] por eso, las actitudes nocionales psicológicas deberían, en principio, ser capaces de captarlas... [pongamos en evidencia, en lo que sigue,] algunos fundamentos adicionales para el escepticismo acerca de los mundos nocionales.

El tema de un mundo nocional, de un mundo *constituido* por la mente o la experiencia del sujeto, ha sido un *leitmotiv* recurrente en la filosofía, al menos desde Descartes. De distintas maneras ha obsesionado al idealismo, al fenomenalismo, al verificacionismo y a la teoría coherentista de la verdad, y a pesar de los palos que típicamente recibe, vuelve a la vida en versiones nuevas y perfeccionadas: en *Ways of Worldmaking* de Goodman (1978) y en la reciente reevaluación del realismo en Putnam (1978), por ejemplo. La ubicuidad del tema no es prueba, en modo alguno, de su solidez, en cualquiera de sus versiones; puede no ser nada más que un error eternamente tentador. En su versión actual, corre cabeza a cabeza con una intuición igualmente convincente acerca de la *referencia*. Si las *actitudes nocionales* van a jugar el rol intermediario que se les asigna, si van a ser la contraparte psicológica de la concepción de Kaplan acerca del *carácter* de una expresión lingüística, entonces debería seguirse que cuando una criatura o sujeto psicológico, con su mundo nocional fijado por su constitución interna, es ubicado en diferentes contextos, en diferentes entornos reales, ello debería determinar diferentes actitudes proposicionales para el sujeto:

actitud nocional + entorno → actitud proposicional.

Esto significa que si yo y mi *Doppelgänger* fuéramos cambiados, instantáneamente (o en todo caso, sin permitir que ocurra ningún cambio de estado interno mientras dure la transición; el intercambio podría tomar un largo tiempo, en tanto yo y mi *Doppelgänger* estuviéramos todo el

25. Kaplan (1968) es explícito: "El rasgo crucial de esta noción [los nombres vívidos de Ralph] es que sólo depende del estado mental presente de Ralph e ignora todos los lazos, sea por semejanza o por génesis, con el mundo actual... Se pretende que abarque los aspectos puramente internos de la individuación" (pág. 201).

tiempo en coma) yo me despertaría con actitudes proposicionales *acerca de las cosas de la Tierra Gemela*, y mi *Doppelgänger* tendría actitudes proposicionales *acerca de las cosas de la Tierra*.[26] Pero esto es altamente anti-intuitivo (para mucha gente, descubro, aunque no para todos). Por ejemplo, tengo muchas creencias y otras actitudes acerca de mi esposa, una persona de la Tierra. Cuando mi *Doppelgänger* se despierta por primera vez en la Tierra después del cambio, y piensa "Me pregunto si Susan habrá preparado ya el café", *seguramente* no tiene pensamientos acerca de *mi esposa*: ¡nunca la ha visto ni ha oído de ella! Sus pensamientos, seguramente, son acerca de *su* Susan, distante años luz, aunque por supuesto, él no sospecha de la distancia. El hecho de que él nunca conocerá la diferencia, ni ningún otro, excepto el Demonio Maligno que llevó a cabo el cambio, es irrelevante; lo que nadie puede *verificar* podría sin embargo ser *verdadero*; sus pensamientos no son acerca de mi esposa, al menos no lo son hasta que no haya tenido algún comercio causal con ella.

Ésta es la esencia de la teoría causal de la referencia (véase, por ejemplo, Kripke, 1972; Evans, 1973; Donnellan, 1966, 1970, 1974) y el experimento mental que la distingue muy bien. Pero las intuiciones provocadas en esas insensatas circunstancias de ciencia ficción son un test pobre.

[...]

TRADUCTORES: Eduardo Barrio y Diana Pérez.
REVISIÓN TÉCNICA: Eduardo Rabossi.

26. Mi *Doppelgänger* no tendría, sin embargo, pensamientos *acerca de mí* cuando pensara "Estoy dormido", y así en más. La referencia del pronombre de primera persona no está afectada por el cambio-de-mundos [*world-switching*], por supuesto (véase Putnam, 1975; Perry, 1977, 1979; Lewis, 1979). Pero uno debe tener cuidado de no convertir este punto en una teoría metafísica de la identidad personal. Consideremos esta variación en un tema de ciencia ficción familiar en filosofía. Nuestra nave espacial se estrella en Marte y uno quiere regresar a la Tierra. Afortunadamente, hay disponible un Teletransportador [*Teleporter*]. Uno se mete en la cabina en Marte y padece un análisis microfísico completo que requiere que uno se disuelva en sus componentes atómicos, por supuesto. El Teletransportador emite la información a la Tierra, donde el receptor, con muchos átomos acumulados a la manera en que una fotocopiadora acumula papel blanco fresco, crea un duplicado exacto de uno, que sale y continúa con su vida en la Tierra con su familia y sus amigos. El Teletransportador, ¿lo "asesinó diseccionándolo", o lo ha transportado a su casa? Cuando el uno-terráqueo, recientemente transportado, dice: "Tuve un terrible accidente en Marte", ¿es verdadero lo que dice? Supongamos que el Teletransportador pueda obtener la información sobre uno sin disolverlo, de tal manera que uno continúa su vida solitaria en Marte. En sus marcas, listos, ya...

REFERENCIAS BIBLIOGRÁFICAS

Aquila, R. E.: (1977) *Intentionality: A Study of Mental Acts*. Penn. State University Press, University Park and London.

Campbell, D.: (1977) "Descriptive Epistemology: Psychological, Sociological and Evolutionary", William James Lectures, Harvard University.

Castañeda, H.-N.: (1966) " 'He': A Study of The Logic of Self-Consciousness", *Ratio* 8.

Castañeda, H.-N.: (1967) "Indicators and Quasi-Indicators", *American Phil. Quarterly* 4.

Castañeda, H.-N.: (1968) "On the Logic of Attributions of Self-Knowledge to Others", *Journal of Philosophy* LXV.

Dennett, D. C.: (1969) *Content and Consciousness*, Londres, Routledge and Kegan Paul.

Dennett, D. C.: (1978) *Brainstorms*, Bradford Books, Montgomery, Vt., and Harvester Press, Sussex.

Dennett, D. C. "True Believers: The Intentional Strategy and Why it Works". [A 1979 Herbert Spencer Lecture, Oxford University.]

Dennett, D. C.: (1987) *The Intentional Stance*, Bradford Books, Montgomery, Vt.

Donnellan, K.: (1966) "Reference and Definite Descriptions", *Philosophical Review* 75.

Donnellan, K.: (1970) "Proper Names and Identifying Descriptions", *Synthèse* 21.

Donnellan, K.: (1974) "Speaking of Nothing", *Philosophical Review* 83.

Evans, G.: (1973) "The Causal Theory of Names", *Aristotelian Society Supplementary Volume* XLVII.

Fodor, J. A.: (1980) "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology", *The Behavioral and Brain Sciences* 3.

Goodman, N.: (1961) "About", *Mind* LXXI.

Goodman, N.: (1978) *Ways of Worldmaking*, Indianapolis, Hackett.

Hampshire, S.: (1975) *Freedom of the Individual*, Princeton University Press.

Hintikka, J.: (1962) *Knowledge and Belief*, Ithaca, Nueva York, Cornell University Press.

Hofstadter, D.: (1979) *Gödel, Escher, Bach: An Eternal Golden Braid*, Basic Books, Nueva York.

James, W.: (1890) *Principles of Psychology*, vol. 1, New York, Dover Publications, 1950.

Kaplan, D.: (1968) "Quantifying In", *Synthèse* 19. [También en *Words and Objections*, comps. D. Davidson y J. Hintikka, Dordrecht, Reidel, 1969.]

Kripke, S.: (1972) "Naming and Necessity", en *Semantics of Natural Language*, comps. D. Davidson y G. Harman, Boston, Reidel.

Lewis, D.: (1978) "Truth in Fiction", *American Phil. Quarterly* 15.

Lewis, D.: (1979) "Attitudes *De Dicto* and *De Se*", *Phil. Review* 87.

Perry, J.: (1977) "Frege on Demonstratives", *Phil. Review* 86.

Perry, J.: (1979) "The Problem of the Essential Indexical", *Noûs* 13.

Putnam, H.: (1975) "The Meaning of 'Meaning' ", en *Language, Mind and Knowledge. Minnesota Studies in the Philosophy of Science* vol., VII, comp. Keith Gunderson, Minneapolis, University of Minnesota P. [Reimpreso en Putnam *Philosophical Papers Vol. II: Mind, Language and Reality*, Cambridge, Cambridge University Press.]

Putnam, H.: (1978) *Meaning and The Moral Sciences*, Londres, Routledge and Kegan Paul.

Schiffer, S.: (1978) "The Basis of Reference", *Erkenntnis* 13.

Searle, J.: (1979) "Referential and Attributive", *The Monist* 62. [También en Searle: *Expression and Meaning*, Cambridge, Cambridge University Press, 1980.]

Stich, S.: (1978) "Autonomous Psychology and the Belief-Desire Thesis", *The Monist* 61.

CAPÍTULO 11

AVISO EN FAVOR DE UNA SEMÁNTICA PARA LA PSICOLOGÍA * (SELECCIÓN)

Ned Block

El concepto de significado es notoriamente vago. Por lo tanto, no debería sorprender que los semánticos [*semanticists*] (los que estudian el significado) hayan tenido en mente propósitos un tanto diferentes y, de tal manera, hayan aguzado el concepto corriente de significado de modos un tanto diferentes. Es un hecho curioso y desafortunado que los semánticos digan poco, habitualmente, acerca de qué aspectos del significado están tratando y cuáles no. Uno tiene pocas pistas acerca de la medida del desacuerdo real entre los programas de investigación "rivales".

Mi propósito aquí es defender un enfoque de la semántica que sea relevante para los fundamentos de la psicología o, mejor, un enfoque para una rama de la psicología, a saber, la ciencia cognitiva. Me expresaré en los términos de algunas de las principales ideas de la ciencia cognitiva; de manera prominente, [en términos de] la teoría representacional de la mente, cuyos aspectos esquematizaré a medida que se tornen relevantes.[1] La doctrina representacionalista, de la que mi argumento depende, sostiene que los pensamientos son entidades estructuradas. Sé que este punto será urticante [*sticking*] para algunos lectores, así que diré un poco más acerca de adónde va a parar esto, y compararé mi posición con otras relacionadas que lo rechazan.

Mi estrategia será comenzar con algunos *desiderata*. Estos *desiderata* varían en muchas dimensiones: cuán centrales son al significado, cuán psicológicamente orientados están, cuán controvertidos son. Argumentaré que un enfoque de la semántica (no los mantendré en suspen-

* "Advertisement for a Semantics for Psychology", *Midwest Studies in Philosophy* (1986), págs. 615-677. Con autorización del autor y de *Midwest Studies in Philosophy*.

1. Algunos buenos esbozos de las ideas de una teoría representacional de la mente se encuentran en Fodor (1981) y Lycan (1981). Un tratamiento más detallado se provee en Pylyshyn (1984).

so: la semántica de rol conceptual [*conceptual role semantics*]) promete tratar tales *desiderata* mejor que otros que conozco. Aunque pienso que mis *desiderata* producen un cuadro coherente de la semántica psicológicamente relevante, no intentan ser obvios preteóricamente, más bien, fueron elegidos para allanar la teoría que tengo en mente. *No* argumentaré que las teorías semánticas que no satisfacen esos *desiderata* son por ello defectuosas; hay problemas distintos —e igualmente legítimos— acerca del significado que una teoría semántica puede proponerse resolver.

El punto de vista que estoy anunciando es una variante del funcionalismo [*functionalism*] familiar en la filosofía de la mente. Sin embargo, no trataré de rebatir las objeciones que se le han formulado (excepto brevemente, y al pasar). Mi apuesta es ésta: ver al funcionalismo desde el punto de vista del significado (en lugar de verlo desde la mentalidad [*mentality*]), prestar atención a su fertilidad y poder en lugar de a su debilidad, nos dará razones para trabajar en sus problemas.

Desiderata

Desideratum 1: Explicar la relación entre el significado y la referencia/verdad. Éste es el menos psicológico de todos mis *desiderata*. Los detalles de lo que tengo en mente se discutirán cuando diga cómo la semántica de rol conceptual promete explicar la relación entre el significado y la verdad.

Desideratum 2: Explicar qué hace significativas a las expresiones significativas. ¿Qué es lo que hay en 'gato', en virtud de lo cual tiene el significado que tiene? ¿Qué diferencia hay entre 'gato' y 'glurg', en virtud de lo cual el primero tiene significado y el último no lo tiene? (Y así sucesivamente para otros tipos de expresiones que no sean palabras.)

Desideratum 3: Explicar el carácter relativo del significado respecto del sistema representacional. Se puede argumentar que este *desideratum* es un caso particular del precedente, pero pienso que vale la pena mencionarlo y discutirlo separadamente. Como todos sabemos, un ítem lingüístico —por ejemplo, un sonido o una expresión lingüística— puede tener distintos significados en distintos lenguajes. Por ejemplo, muchos ítemes del vocabulario tienen significados diferentes en los dialectos del inglés que se hablan en Estados Unidos e Inglaterra, como *trailer* y *bathroom*.

Pero la significatividad [*significance*] del carácter relativo del significado respecto del sistema de representación resulta más profunda de lo que estos ejemplos sugieren. Una manera de ver esto es advertir que todas las *categorías* semánticas (y sintácticas) son relativas al sistema de representación. Las manchas de tinta que funcionan como una figura [*picture*] en su tribu, pueden funcionar como una palabra en la mía. Más aún, dentro de la categoría de las figuras, las representaciones se comprenden de maneras diferentes en diferentes culturas.[2] Finalmente, la categoría sintáctica es relativa de manera similar. La letra manuscrita, por ejemplo, varía en diferentes sistemas escolares. Tal vez las manchas de tinta que se ven como una 'A' en Edimburgo se vean como una 'H' en Chicago. ¿Hay alguna explicación común de la relatividad del sistema representacional tanto a las categorías semánticas como a las sintácticas?

Desideratum 4: Explicar la composicionalidad. El significado de una oración es en algún sentido una función de los significados de las palabras que hay en ella (más la sintaxis de la oración). ¿Cuál es, exactamente, la relación entre los valores semánticos [*semantic values*] de las oraciones y de las palabras? ¿Es uno más básico que el otro? Cuando uno ha dado respuesta a estas preguntas surge otra pregunta, a saber, ¿por qué resulta que el valor semántico de una oración tiene la relación que tiene con el valor semántico de sus partes?

Desideratum 5: Concordar con una explicación de la relación entre el significado y la mente/cerebro. ¿Por qué uno debería esperar (o al menos desear) que una teoría *semántica* concuerde con una explicación de la relación entre el significado y la mente o el cerebro? Porque sería sorprendente que la naturaleza del significado (lo que el significado *es*) sea totalmente irrelevante para explicar qué es entender [*grasp*] o comprender significados y cómo [es que] entender significados puede tener efectos físicos. Al menos, uno puede imaginar diferencias entre *x* e *y* que hagan a una diferencia entre lo que es entender *x* e *y*. Por ejemplo, comprender *x* puede requerir destrezas o habilidades de reconocimiento, mientras que comprender *y* puede requerir sólo conocimiento proposicional.

Dije "mente *o* cerebro", pero de hecho me centraré en el cerebro.

2. Véase Block (1983) para una discusión de esta distinción y para referencias de la literatura acerca de este tópico.

Y al discutir este asunto adoptaré simplemente una forma de materialismo (la tesis de la identidad de "casos" [*"token" identity thesis*]: cada acaecimiento [*occurrence*] mental particular es un acaecimiento físico).

¿Qué se supone que necesita ser explicado en la relación entre el significado y el cerebro? Bueno, una pregunta obvia es: ¿en qué consiste que el cerebro entienda significados, y cómo es que entender significados por parte del cerebro tiene efectos en el mundo? Los significados son (al menos aparentemente) objetos abstractos no-físicos. Y la relación entre el cerebro y los significados que el cerebro entiende no parece ser como la relación entre una varilla de metal y el número de grados Celsius que es su temperatura, un caso en el que hay propuestas acerca de cómo un cambio en el valor de la temperatura puede causar, digamos, la expansión de la varilla (véase Field, 1980). No obstante, la diferencia entre un cerebro que entiende un cierto significado y un cerebro que no lo hace, cuenta como una diferencia en las propiedades causales de ese cerebro. Un cerebro que entiende el significado de 'transmogrify' puede hacer que su poseedor gane en un show de preguntas y respuestas, transportándolos a ambos a un hotel en Catskills. Necesitamos una explicación de cómo esa relación entre un cerebro y un significado puede producir una diferencia causal.

Desideratum 6: Aclarar la relación entre el significado autónomo y el heredado [*autonomous and inherited meaning*]. Si hay representaciones en el cerebro, como sostiene la teoría representacional de la mente, entonces hay que trazar una distinción obvia entre estas y otras representaciones; por ejemplo, las representaciones de esta página (Searle, 1980; Haugeland, 1980). Las representaciones de esta página tienen que leerse o escucharse para ser comprendidas, pero ese no es el caso para las representaciones en un cerebro. Las representaciones de esta página requieren ser *traducidas* para ser comprendidas, o al menos [requieren] una *trasliteración* [*transliteration*] al lenguaje del pensamiento; las representaciones en el cerebro (algunas de ellas, en todo caso) no requieren traducción o trasliteración. Digamos que las representaciones que no requieren traducción o trasliteración tienen un significado *autónomo*, mientras que las que requieren traducción o trasliteración tienen un significado *heredado*.

Diferentes puntos de vista acerca del significado tienen consecuencias muy diferentes para el tema de qué podría decir una teoría semántica acerca de ambos tipos de significado. Según el punto de vista de Searle, por ejemplo, lo máximo que una teoría semántica podría decir

sobre este asunto es dar una explicación de cómo el significado heredado (significado *relativo-al-observador* [*observer-relative meaning*], en su terminología) se hereda del significado autónomo (*significado intrínseco* [*intrinsic meaning*], en su terminología). Explicar el significado autónomo en sí mismo, según Searle, cae simplemente fuera del alcance de la semántica. Lo máximo que podemos decir para dar una explicación del significado autónomo es que proviene de los poderes causales del cerebro humano, y que provendría de cualquier otro objeto (por ejemplo, una máquina) que tuviera "poderes causales equivalentes".

A pesar de la variedad de puntos de vista acerca de este asunto, hay unas pocas preguntas cuyo interés debería hallar acuerdo entre quienes aceptan la distinción entre significado autónomo y heredado, como punto de partida. Las preguntas principales son: ¿qué son el significado autónomo y el heredado?; ¿cuál es la relación entre el significado autónomo y el heredado? Por ejemplo, ¿son dos tipos diferentes de significado, ninguno de los cuales es derivado del otro, o reducible a él?[3]

Un problema relacionado es cómo una representación con significado autónomo puede significar lo mismo que una representación con significado heredado. Muchos filósofos menospreciarán tal pregunta a causa del escepticismo acerca de la sinonimia. Pero no está claro que quienes la aceptan queden atrapados en las arenas movedizas quineanas. Eso depende de que la noción de significado usada en la ciencia cognitiva tenga que acarrear consigo un compromiso con *verdades* acerca del significado [*truths of meaning*] y, por lo tanto, un compromiso con la verdad a priori.[4]

Desideratum 7: Explicar las conexiones entre conocer, aprender y usar una expresión, y el significado de la expresión. Obviamente, hay una conexión estrecha entre *el significado de una palabra*, de un lado, y *lo que conocemos cuando conocemos o comprendemos una palabra* y *lo que aprendemos cuando aprendemos una palabra*, del otro. Sin duda que es intuitivamente plausible que estas descripciones italizadas tengan el

3. Espero que mi terminología "heredado/autónomo" no haga parecer triviales estos problemas.

4. Es el compromiso con una verdad *a priori* (con lo que quiero decir verdades para las cuales no hay posibilidad cognitiva [*epistemic*] de refutación) lo que realmente causa problemas a los amigos de la analiticidad, no nuestra ineptitud para proponer condiciones de identidad para el significado. Después de todo, nadie ha propuesto nunca condiciones de identidad satisfactorias para las personas o los barcos.

mismo referente (aunque sería un error adherir dogmáticamente a esta intuición preteórica).

Más aún, quien ha aprendido una expresión (y por lo tanto la conoce) automáticamente tiene la capacidad de usarla correctamente; además la evidencia del uso correcto es evidencia del conocimiento del significado. Una teoría del conocimiento psicológicamente relevante ha de iluminar las conexiones entre conocimiento/comprensión/aprendizaje y uso, de un lado, y el significado, del otro.

Desideratum 8: Explicar por qué diferentes aspectos del significado son relevantes, de maneras diferentes, para determinar la referencia y la explicación psicológica. Se pueden distinguir dos aspectos del significado que son relevantes de maneras muy diferentes para la explicación psicológica. Un tipo de casos involucra deícticos [*indexicals*], por ejemplo:

(1) Yo corro el riesgo de ser atropellado.
(2) Ned Block corre el riesgo de ser atropellado.

Considérese la diferencia entre las creencias que yo expresaría al proferir [*by uttering*] (1), en comparación con (2). No se puede garantizar que creer (2) tenga el mismo efecto protector sobre mi conducta, que creer (1), dado que puedo no saber que soy Ned Block (puedo pensar que soy Napoleón).[5] Así, hay una diferencia importante entre (1) y (2) con respecto a la causación (y por tanto con respecto a la explicación causal) de la conducta.

Esta observación motiva una manera familiar de pensar acerca del significado y el contenido de las creencias [*belief content*], de acuerdo con la cual cuando usted y yo tenemos creencias que se expresan mediante nuestras (respectivas) emisiones de (1), tenemos creencias con el mismo contenido. Ésta es la manera de individuar [*individuating*] según la cual dos lunáticos que dicen "Soy Napoleón" tienen la *misma errónea ilusión* [*delusion*]. Correspondiendo a esta manera de individuar el contenido de las creencias, tenemos una manera de individuar los significados, según la cual los significados de las oraciones-caso [*sentence tokens*] de los dos lunáticos, son el mismo [significado]. Ésta es la manera de individuar los significados de los casos dirigida a [*is geared towards*] las oraciones-tipo [*sentence types*], y por lo tanto parece muy natural para la lingüística, dado que hace del significado de la oración

5. Perry (1977, 1979); Kaplan (inédito).

una función de los significados de las palabras de la oración (más la sintaxis). Nótese que según esta manera de individuar, mis emisiones de (1) y (2) tienen *diferentes* significados y expresan de manera estándar creencias con *diferentes* contenidos. Además, esta manera de individuar es natural para la lingüística, dado que ningún diccionario razonable daría la misma entrada a "Yo" y a "Ned Block".

Sin embargo, (1), dicho por mí, y (2), expresan la misma proposición de acuerdo con una manera familiar de individuar proposiciones. En un sentido familiar de 'significado', de acuerdo con el cual dos oraciones-caso tienen el mismo significado sólo si expresan la misma proposición, (1), dicha por mí, y (2) tienen el mismo significado. Si individuamos los contenidos de las creencias tal como individuamos las proposiciones creídas, la creencia que expreso mediante (1) tendría el mismo contenido que la que expreso mediante (2). Más aún, la creencia que expreso mediante (1) tendría un contenido diferente de la creencia que usted expresa mediante (1); de la misma manera, el significado de mi proferencia de (1) sería diferente del de su proferencia de (1).

Llámese al primer esquema de individuación, individuación *estrecha* [*narrow*], y al último, individuación *amplia* [*wide*] (*cf.* la distinción diferente que traza Kaplan entre carácter [*character*] y contenido). La individuación amplia agrupa las oraciones-caso si atribuyen las mismas propiedades a los mismos individuos, mientras que la individuación estrecha agrupa las oraciones-caso si atribuyen las mismas propiedades usando las mismas descripciones de los individuos, con independencia de si los individuos referidos son los mismos. Con otras palabras, la individuación estrecha hace caso omiso de la pregunta de (esto es, ignora) si los mismos individuos están involucrados y depende, en cambio, de cómo los individuos son referidos.[6] (Nótese que la pregunta de cómo los individuos son referidos es muy diferente de la pregunta de cómo el que hace la referencia piensa al referente. Por ejemplo, dos usos de (1) tienen el mismo significado estrecho (en mi sentido de la frase) aun si un usuario piensa que es Napoleón mientras que otro piensa que es Wittgenstein.)

Uno puede pensar a la individuación estrecha y a la amplia como especificando aspectos diferentes del significado: el significado estrecho y el amplio. (No estoy diciendo que el significado estrecho y el amplio sean *tipos* de significado, sino sólo aspectos o tal vez *determinantes*

6. Una variante natural de la noción de individuación estrecha que he descripto requeriría además que las mismas propiedades sean atribuidas de la misma manera.

[*determinants*] del significado.) El significado estrecho está "en la cabeza", en el sentido de esta frase que indica la superveniencia [*supervenience*] a la constitución física,[7] y el significado estrecho captura el aspecto semántico de lo que tienen en común las proferencias de (por ejemplo) (1) por parte de diferentes personas. El significado amplio, por contraste, depende de qué individuos fuera de la cabeza son referidos, así, el significado amplio no está "en la cabeza". El tipo de individuación que da origen al significado estrecho también da origen al concepto correspondiente de contenido estrecho de las creencias. Dos proferencias tienen el mismo significado estrecho sólo si las creencias que expresan tienen el mismo contenido estrecho.

Nótese que a pesar de la terminología engañosa, el significado amplio no *incluye* al significado estrecho. Las proferencias de (1) (hechas por mí) y (2) tienen el mismo significado amplio pero no el mismo significado estrecho.[8]

El significado/contenido estrecho y el significado/contenido amplio son relevantes para la explicación psicológica, de maneras muy diferentes. Así, el significado estrecho de la oración creída es más informativo respecto del estado mental de quien cree [*believer*]. De tal modo, el significado estrecho (y el contenido estrecho) es más apropiado para predecir y explicar lo que alguien decide o hace, en tanto la información sobre el mundo externo sea ignorada. Así, si usted y yo tenemos una creencia que se expresaría por medio de (1), uno podría explicar y predecir nuestras repentinas miradas a los vehículos que nos rodean, y nuestras decisiones (respectivas) de saltar a un costado. Los significados amplios son menos apropiados para este tipo de predicción y explicación, porque "dejan afuera" información acerca de la manera en que uno se refiere a sí mismo. Dado que el significado amplio de (1), dicho por mí, y de (2), es el mismo, si se le dice que creo una oración con ese significado amplio (esto es, el significado amplio común a mi [1] y [2]),

7. Nótese que la pretensión de que el significado estrecho está en la cabeza, en este sentido, no es incompatible con la idea de que lo que es para una palabra tener un cierto significado estrecho es, para ella, expresar un concepto, donde los conceptos se toman como objetos abstractos, no localizables en el espacio ni en el tiempo; en este sentido, "en la cabeza" no es una frase apta.

8. Por supuesto, uno podría definir una noción referencial del significado que incluya el significado estrecho y que por lo tanto sea más merecedor de ser llamado "amplio". Esto también resultaría en un tratamiento más intuitivo de la referencia vacua. Dado que el uso principal que haré de esta noción de significado amplio es iluminar el significado estrecho, me atendré a la definición simple que he introducido.

usted sabe que yo creo que algo —yo, en este caso, aunque no se le informa que yo sé que se trata de mí— está en peligro de ser atropellado. Así, la información es omitida, dado que no se le informa cómo concibo la cosa que está en peligro. Por otra parte, usted sí sabe que yo creo que algo está en peligro, de modo que usted tiene *alguna* información sobre mi estado mental.

Podría parecer, a partir de lo que acabo de decir, que el significado estrecho incluye todo lo que es relevante para la explicación psicológica que el significado amplio incluye, y aún más. Pero en un sentido el significado amplio puede ser más útil para la predicción: en la medida en que haya relaciones nomológicas entre el mundo y lo que la gente piensa y hace, el significado amplio permitirá predecir lo que ella piense y haga sin contar con información acerca de cómo ve las cosas. Supóngase, por ejemplo, que la gente tiende a evitar los espacios abiertos, sin importar cómo se describa a sí misma esos espacios. Entonces, al saber que Fred está eligiendo entre ir por un espacio abierto o por una calle de la ciudad, uno estaría en posición de predecir la elección de Fred, aunque uno no sepa si Fred describe para sí al espacio abierto como 'esto' o como cierta plaza.

El significado estrecho tiene otro tipo de significación [*import*] teórica: determina una función de expresiones y contextos de emisión a referentes y valores de verdad.[9] Cuando usted y yo emitimos 'Yo' en (1), hay algo que compartimos, algún aspecto semántico de la palabra 'Yo' que en su contexto mapea [*maps*] su [oración-] caso en usted [*onto you*], y que en mi contexto mapea mi [oración-]caso en mí.

Permítaseme protegerme de algunos malentendidos. Primero, como ya he indicado, el significado estrecho de 'Yo' no incluye la propia concepción de uno mismo. Segundo, aunque he dicho que hay un aspecto semántico compartido de 'Yo' que es relevante para la explicación de la conducta y un aspecto semántico compartido para determinar una función de contexto a referente, no sugiero que esos aspectos semánticos compartidos sean exactamente el mismo. Es una pregunta pendiente [*open question*] si son lo mismo, y por lo tanto si el 'significado estrecho', como estoy usando el término, señala [*picks out*] una sola cosa. En la teoría por la que abogo, el aspecto semántico que determina la función de contexto a referente (y el valor de verdad) se vuelve una *parte* del aspecto semántico que desempeña un rol en la explicación de la conducta. Así, este último aspecto semántico hace *ambas* tareas. Por lo tanto,

9. Véase Loar (1982), 279; White (1982); y Fodor (1985).

usaré 'significado/contenido estrecho' como refiriendo únicamente al aspecto semántico más inclusivo. Quiero señalar, sin embargo, que esta manera de hablar conlleva un fuerte compromiso teórico. Finalmente, la distinción estrecho/amplio, tal como la describí, se aplica a ejemplares, no a tipos. Sin embargo, hay una extensión obvia a los tipos (no-deícticos).

Haré una pausa ahora para decir que las consideraciones presentadas hasta aquí en esta sección tienen que ver con una semántica para la psicología. En primer lugar, una semántica para la psicología debería tener algo que decir acerca de la proveniencia de la distinción entre el significado estrecho y el amplio e, idealmente, debería dar cuenta de qué son los dos aspectos del significado. Segundo, la teoría debe decir por qué los significados estrechos y amplios son claramente relevantes para la explicación y predicción de los hechos psicológicos (incluida la conducta). Tercero, la teoría ha de dar una explicación del significado estrecho que explique cómo es que determina una función que va del contexto de emisión a la referencia y al valor de verdad.

He hablado hasta aquí del significado de las oraciones con deícticos, pero los puntos que he formulado pueden extenderse a los nombres, y, más controvertidamente, a los términos de clases naturales. Considérese a Teen (de la Tierra) y su gemela en la Tierra Gemela, Teen_{tg}. Las dos son duplicados partícula por partícula que han tenido exactamente la misma historia de estimulaciones de superficie [*surface stimulations*]. En algunas de las diferentes versiones del relato tenemos que imaginar varias diferencias en los mundos que las rodean, fuera de la esfera de lo que las ha impactado [*impinged*]. Supongamos ahora que su medio ambiente es exactamente idéntico, excepto, por supuesto, que los individuos de los dos mundos son distintos: el héroe de Teen es Michael Jackson, mientras que el héroe de Teen_{tg} es un personaje distinto pero indistinguible (excepto espaciotemporalmente). Teen y Teen_{tg}, cada una de ellas, tiene un pensamiento que podrían expresar con:

(3) Michael Jackson se contonea.

Podemos distinguir nuevamente entre dos maneras de individuar el contenido de los pensamientos y también el significado de las oraciones pensadas. En uno, el esquema estrecho, podemos decir que Teen y Teen_{tg} tienen el mismo pensamiento, y podemos decir que emiten oraciones con el mismo significado. Si ambas dijeran sinceramente "Michael Jackson tiene poderes sobrenaturales", compartirían la misma erró-

nea ilusión. Éste es el significado estrecho y el contenido estrecho. Alternativamente, podemos ver los significados y los contenidos del pensamiento como distintos en virtud del hecho de que Teen se refiere a Michael Jackson y $Teen_{tg}$ se refiere a Michael $Jackson_{tg}$. Éste es el significado y el contenido amplio.

Esto ilustra [un caso de] el mismo significado y contenido estrechos, pero distinto significado y contenido amplios [*same narrow/different wide meaning and content*]. El caso de un mismo significado estrecho, pero distinto significado amplio [*same wide/different narrow meaning*] (el caso anterior análogo a [1] y [2] proferido por la misma persona) se ilustra con 'Cicerón ora [*orates*]' y 'Tulio ora'. En estos casos de nombres [*name cases*] los principios de individuación son los mismos que en los casos de deícticos, aunque su motivación es más débil, en un aspecto, porque es discutible que los nombres *tengan* significado. Además, la conexión nomológica entre los nombres y la conducta no es tan simple como la que hay entre 'Yo' y la conducta.

Hay dos hechos básicos en los cuales está basada la distinción estrecho/amplio. [El primero] es que [la manera] como uno represente algo a lo que uno refiera, puede afectar nuestros estados psicológicos y nuestra conducta. Así, si usted sabe que Cicerón ora y no sabe que Cicerón = Tulio, no está en condiciones de apelar al hecho de que Tulio ora. El segundo hecho básico es que hay más [cosas relevantes] para la semántica que lo que está "en la cabeza". Los contenidos de la cabeza de una persona que afirma (3), conjuntamente con el hecho de que Michael Jackson se contonea, *no son suficientes para determinar si (3) es verdadera o falsa*, dado que el valor de verdad depende también de quién es referido con 'Michael Jackson'. Imaginemos que aunque Michael Jackson se contonea de manera excelente, su gemelo no puede contonearse; y los contoneos adscriptos a su gemelo por las adolescentes de la Tierra Gemela son efectivamente realizados por un doble. Entonces, las proferencias [*utterances*] de (3) en la Tierra Gemela difieren en valor de verdad de las proferencias de (3) en la Tierra, a pesar de que no hay diferencias relevantes en las cabezas de las adolescentes de los dos planetas, y a pesar de ser un hecho que Michael Jackson se contonea tanto en la Tierra como en la Tierra Gemela. (Si esto le parece misterioso, nótese que en la última oración, yo usé 'Michael Jackson' como se usa en mi comunidad lingüística —¿podría yo hablar el lenguaje de algún otro?— y la comunidad lingüística de la Tierra Gemela usa la misma expresión para referir a una persona diferente.) *Dado que el valor de verdad de una oración está determinado por la totalidad de los hechos*

semánticos, más los hechos relevantes acerca del mundo, hay más, respecto de la totalidad de los hechos semánticos acerca de una oración, que lo que hay en la cabeza del hablante. Los hechos semánticos "extra" son acerca de aquello a lo que los términos referenciales de la oración refieren.[10] Pero aun cuando haya diferencias semánticas entre las emisiones de (3) y pensar en (3) por parte de Teen y de Teen_{tg}, hay similitudes importantes también —y éste es el punto principal de la presente sección— que dan lugar a las nociones de los aspectos del contenido y del significado (contenido y significado estrechos) *que son compartidos por Teen y* $\textit{Teen}_{tg}$, que explican las similitudes en (por ejemplo) la vida de fantasía [*fantasy life*] y la conducta de compra de entradas [*ticket-buying behavior*] y que determinan la función [que va] de diferentes contextos a sus diferentes referentes.

Como en el caso del deíctico, el significado y el contenido amplios no son apropiados para explicar el cambio de estado mental y la conducta. El significado amplio de 'El agua es húmeda' (en español, no en español gemelo) es el mismo que el de 'H_2O es húmeda', a pesar de los efectos potencialmente diferentes que creer en esas oraciones tenga en los estados mentales y en la conducta. Más aún, como revela el ejemplo de Pierre dado por Kripke (Kripke, 1979), si la concepción que uno tiene de la traducción es expresamente referencial (permitiendo que 'London' se traduzca por 'Londres' en los contextos de creencia), uno se enfrenta con situaciones en las cuales se ve forzado a adscribir creencias contradictorias que no son atribuibles a quien cree.[11] Además, lo que comparten Teen y Teen_{tg} determina también que una se refiera a Michael Jackson, mientras que la otra se refiere al gemelo de Michael Jackson. Lo compartido determina una función de contexto a referencia. Si Teen hubiera sido criada en la Tierra Gemela, hubiera sido la misma, molécula por molécula, a como efectivamente es (ignorando la indeterminación cuántica), pero su ejemplar de 'Michael Jackson' habría referido al gemelo de Michael Jackson.[12]

El lector puede preguntarse por qué he dedicado un espacio tan

10. Cf. Field (1977).

11. Ésta es una lectura controvertida de la "lección" del acertijo [*puzzle*] de Kripke. No tengo espacio aquí para describir ni la solución del acertijo, ni la de la semántica de rol conceptual.

12. White (1982) trata de *definir* una noción de significado estrecho usando tales contrafácticos. Pero esto parece mal orientado, dado que hay algo compartido por los gemelos *en virtud de lo cual* los contrafácticos son verdaderos, y esto parece un candidato mejor para el significado estrecho.

extenso a este *desideratum* (acerca de la distinción estrecho/amplio). (¡Y todavía no he terminado!). La versión de la semántica de rol conceptual que defenderé caracteriza el significado *estrecho* en términos de rol conceptual. Hay otra versión (Harman, 1982) que no tiene relación con el contenido y el significado estrechos. Los roles conceptuales de Harman involucran interacciones perceptuales y conductuales con lo que se ve y se manipula, esto es, objetos en el mundo, mientras que mis roles conceptuales se detienen en la piel. (Así, si a usted no le gusta que se hable de estrecho, todavía puede apreciar los *desiderata* anteriores como motivando una versión harmaniana de la semántica de rol conceptual). Prefiero mi versión, y estoy tratando de explicitar parte de la motivación en favor de ella.[13] (Diré más sobre la alternativa de Harman luego).

Considérese la historia original de la Tierra Gemela de Putnam. Mi *Doppelgänger* (nuevamente, un duplicado físico)[14] usa 'agua' para referirse a XYZ. Supóngase, siguiendo a Putnam, que XYZ *no* es un tipo de agua. Más aún, podemos agregar a la historia ideas desarrolladas por Burge (Burge, 1979) que muestran las diferencias en la manera en que el uso de las palabras en nuestras diferentes comunidades lingüísticas [*language communities*] puede determinar diferencias en el significado de nuestras palabras, aun cuando no resulten diferentes en los estímulos que afectan nuestra superficie. Supongamos que mi gemelo y yo nos decimos a nosotros mismos:

> Mis pantalones están ardiendo. Pero por suerte, estoy parado delante de una pileta de natación llena de agua. El agua, a Dios gracias, apaga el fuego.

Si Burge y Putnam están en lo correcto (y me inclino a concordar con ellos), hay diferencias semánticas sustanciales entre los significados y los contenidos de pensamiento de mi gemelo y los míos, debido a las diferencias en el medio ambiente físico y social. Sin embargo —y aquí está, nuevamente, la idea crucial que se encuentra detrás de mi defensa del significado y el contenido estrechos—, ***hay algún aspecto del significado en común entre lo que él dice y lo que yo digo (o al menos un***

13. Véase McGinn (1982), especialmente págs. 211-16, para argumentos que van de la naturaleza de la representación al contenido y al significado estrechos.

14. Ignórese el problema de que dado que estamos constituidos mayoritariamente por agua, mi gemelo y yo no podemos ser duplicados —algunos parches [*fixes*] han sido propuestos para esto por parte de Putnam y Burge.

determinante parcial común del significado), y ese aspecto semántico común de lo que decimos provee parte de la explicación común de por qué ambos saltamos a las respectivas piletas. Y si las ideas actuales acerca de la teoría representacional de la mente son correctas, el significado y el contenido estrechos serán útiles para establecer generalizaciones nomológicas relativas al pensamiento, a la decisión y a la acción.

Más aún, si mi gemelo hubiera crecido en mi contexto, su ejemplar de 'agua' referiría a H_2O en lugar de a XYZ. Como en casos anteriores, parece haber algún aspecto semántico común de nuestros términos que opera en mi caso para establecer una correspondencia entre mi contexto y H_2O, y [opera en] su caso para hacer corresponder su contexto con XYZ.

El lector puede haber advertido cómo pasé a la extensión natural de lo que describí como la distinción estrecho/amplio, a partir [de la distinción] caso/tipo. Dado que 'Cicerón' y 'Tulio' son usados de forma estándar para referir a la misma persona, podemos considerar a las oraciones-tipo 'Cicerón ora' y 'Tulio ora' como poseyendo el mismo significado amplio. Lo mismo vale para 'agua' (tal como se usa en español, en oposición al español gemelo) y para 'H_2O'.

Digamos que una adscripción de actitud proposicional o de significado es individualista si es superveniente al estado físico del cuerpo del individuo, en donde el estado físico se especifica no-intencionalmente y con independencia de las condiciones físicas y sociales que se dan fuera del cuerpo.[15] Creo que hay un esquema [*scheme*] individualista importante para la individuación de las creencias, los contenidos de creencias y el significado de las oraciones en las que se cree. Hay un fuerte elemento de individuación individualista en nuestro pensamiento ordinario, pero su hogar natural se encuentra en el pensamiento científico acerca de la mente, especialmente en la ciencia cognitiva contemporánea. También concuerdo con Burge y Putnam en que hay un esquema de individuación importante no-individualista en el pensamiento ordinario. Hasta aquí no hay incompatibilidad.

Pero Putnam, Burge y otros han argumentado también en contra de la individuación individualista. La conclusión de Putnam (1983) está basada en el argumento de que es imposible proponer condiciones de identidad al contenido o al significado considerados de manera individualista. No tengo condiciones de identidad para ofrecer, pero me inclino a ver esto no como un obstáculo insuperable sino como un problema

15. Burge (1979).

a ser disuelto mediante una construcción teórica. Mi conjetura es que una concepción científica del significado debería dejar a un lado la cruda dicotomía mismo/diferente significado, en favor de una gradiente multidimensional de similaridad de significado.[16] Después de todo, [es con] la sustitución de una dicotomía por un continuo que la teoría bayesiana de la decisión evita una cantidad de dificultades —por ejemplo, la paradoja del prefacio— yendo de los crudos casilleros *cree/no cree* a grados de creencia.[17]

Burge (1984) argumenta principalmente contra el "panindividualismo" ["*pan-individualism*"], la pretensión de que *todas* las individuaciones de actitudes proposicionales en psicología sean individualistas. Sin embargo, no estoy defendiendo esa doctrina sino sólo la pretensión más limitada de que hay una importante corriente de individuación individualista en psicología (y en el discurso de sentido común). Burge tiene dudas acerca de esto también, pero el asunto puede ser resuelto sólo por medio de una discusión detallada de la práctica psicológica.

Déjeseme formular sólo una consideración. La psicología se ocupa a menudo de explicar las diferencias psicológicas. La medida de estas diferencias es la *varianza* [*variance*].[18] Por ejemplo, la varianza en la inteligencia y otros atributos y estados psicológicos se atribuye a diferencias en los genes y en el medio ambiente (y a las interacciones de varios tipos entre esos factores causales). Supóngase que llenamos un autobús de excursión con viajeros, la mitad de la Tierra Gemela y la otra mitad de la Tierra. Los terráqueos creen que el agua es húmeda y prefieren beber agua en vez de gasolina, mientras que los habitantes de la Tierra Gemela no tienen esas actitudes proposicionales (porque cuando piensan acerca de lo que ellos llaman 'agua', no piensan acerca del agua, no tienen un término que refiera al agua). Supongamos que los terráqueos y los habitantes de la Tierra Gemela no difieren de manera relevante en los genes o en la estimulación de superficie que ha impactado en sus cuerpos a lo largo de toda su vida. Por tanto, en esa población,

16. En realidad, mi posición es que tal gradiente multidimensional es necesaria para un significado estrecho de pura raza, pero no para la *parte* del significado estrecho responsable de mapear contextos en referentes y condiciones de verdad.

17. Véase Horwich (1982). Aquí está la paradoja del prefacio: escribo un libro en cuyas oraciones creo; no obstante, estoy seguro de que siendo humano, he afirmado al menos una falsedad. Contradicción. Solución: tengo un alto grado de creencia respecto de cada oración del libro, pero eso es compatible con un alto grado de creencia en la falsedad de su conjunción.

18. La varianza es la desviación cuadrática media con respecto a la media.

las diferencias en las actitudes proposicionales no pueden atribuirse al medio ambiente (en el sentido de la estimulación de superficie) y a los genes (y sus interacciones): las diferencias en las actitudes acerca del agua [*water attitudes*] se deben a algo que no tiene nada que ver con las diferencias en los genes o en las estimulaciones de superficie que han afectado [*affected*] a esa gente. Un análisis de la varianza debería atribuir un componente grande de varianza a las diferencias en un factor que no causa ninguna diferencia en las proteínas, conexiones sinápticas o cualquier otro rasgo fisicoquímico del cuerpo, tal como lo hacen las diferencias en los genes y en las estimulaciones de superficie. Esto contaría como un tipo de acción a distancia, e iría claramente en contra de la metodología de atribución de varianza. (Nótese que esto podría haber sido formulado en términos del punto formulado por Burge acerca de la naturaleza social del significado, en lugar de [formularlo en términos de] la Tierra Gemela.)

Acabo de argumentar a favor de la individuación individualista de los estados de actitud proposicional, por ejemplo, las creencias. Pero hay un hiato [*gap*] entre individuar a las creencias de manera individualista e individuar de tal manera a los *contenidos* de las creencias. Uno podría sostener que cuando la creencia se individua individualísticamente, se tiene, aun, una creencia de un tipo extraño, pero que el contenido, individuado de manera individualista, es como un presidente que ha sido depuesto: no es más un presidente (cf. Stich, 1983). Propongo salvar el hiato como sigue.

En ciertos tipos de casos, toda vez que tenemos una relación, tenemos propiedades individualistas [*individualistic properties*] de las entidades relacionadas que, podría decirse, fundamentan la relación. Si *x* golpea a *y*, *y* tiene un cierto tipo de cambio consecuente en la superficie corporal, tal vez la nariz aplastada, y *x* tiene la propiedad de, digamos, mover el puño hacia adelante. Por supuesto, la misma propiedad individualista puede subyacer a muchas propiedades relacionales diferentes, y algunas relaciones no dependen, notoriamente, de las propiedades individualistas, por ejemplo, 'a la izquierda de'. Cuando el contenido se individúa *no*-individualísticamente, se individúa con respecto a las relaciones con el mundo (como en el caso de la Tierra Gemela) y a la práctica social (como en el ejemplo de Burge acerca de la artritis).[19] Hay un

19. Burge (1979). Burge construye casos en los cuales un hombre tiene leves malentendidos respecto de cómo se usa una palabra (por ejemplo, piensa que uno puede tener artritis en el muslo). Argumenta luego, persuasivamente, que un *Doppelgänger* de

aspecto no-relacional del contenido de la actitud proposicional, el aspecto "dentro de la cabeza" ["*inside the head*"], que corresponde al contenido, de la misma manera que mover el puño hacia adelante corresponde a golpear. El aspecto no-relacional del contenido es lo que yo llamo el contenido estrecho. Pero, ¿es el contenido estrecho realmente contenido?[20]

Encuentro mucha hostilidad entre los filósofos respecto de las ideas de contenido estrecho y de significado estrecho. Hay muchas razones para tal resistencia, razones que acepto como temas de una controversia genuina y sobre las cuales no tengo una confianza absoluta en lo que hace a mi posición. Pero la inquietud mencionada parece estar fuera de lugar, al menos como crítica de la semántica de rol conceptual. La crítica es que yo he supuesto erróneamente que el aspecto del significado o del contenido que está dentro de la cabeza es genuinamente *semántico*. Jerry Fodor me acusó una vez de [cometer] la "falacia de sustracción" ["*fallacy of substraction*"], esto es, de suponer que si uno toma el significado o el contenido y *sustrae* su relación con el mundo y su aspecto social, lo que queda es algo semántico.

Por supuesto que *hay* algo que es la falacia de sustracción. Si se sustrae de la rojez la propiedad de ser coloreado, no se obtiene una rojez sin coloración. Pero respecto a la semántica de rol conceptual, el tema es meramente verbal. Nada en mi posición requiere que yo vea el significado estrecho y el contenido estrecho como (respectivamente) *tipos* de significado y contenido. Como se señaló antes, los veo como aspectos o *determinantes* [*determinants*] del significado y el contenido. Todo lo que se requiere para mi posición es que lo que llamo significado estrecho sea un rasgo claro [*distinct*] del lenguaje, cuya caracterización contribuye de manera importante a una teoría total del significado (por ejemplo, como se indicó en mis *desiderata*). Lo mismo vale para el contenido estrecho.

¿Estoy concediendo que la semántica de rol conceptual no es realmente parte de la *semántica*? Lo primero que cabe decir respecto de esta pregunta es que tiene una importancia intelectual muy menor. Es una disputa acerca del límite entre disciplinas; como muchas de tales dis-

---

ese hombre en una comunidad lingüística en la cual 'artritis' se usa de manera estándar incluyendo inflamaciones reumáticas de huesos tales como el fémur, no debería ser visto como significando con 'artritis' lo que nosotros y nuestro hombre significamos con esta palabra.

20. LePore y Loewer (1985) parecen objetar de esta manera la semántica de rol conceptual de dos factores.

putas, sólo puede resolverse por medio de un tipo de filosofía del lenguaje ordinario aplicada a términos técnicos como 'semántica' (o peor aún, por las administraciones universitarias). Cabe a la filosofía del lenguaje ordinario el análisis de los conceptos que desempeñan un rol central en el pensamiento humano ordinario, pero la aplicación de esas técnicas a los términos técnicos, en donde la estipulación está a la orden del día, no es muy iluminadora. Sin embargo, estoy tan dispuesto a objetar [*quibble*] como cualquiera. La aplicación correcta de los términos [técnicos] depende en gran medida de los desarrollos de las disciplinas correspondientes. A menudo las ideas preteóricas acerca del dominio de una disciplina se dejan a un lado. Si el significado realmente se descompone en dos factores, entonces el estudio de la naturaleza de esos dos factores pertenece al dominio de la semántica, aun si uno o ambos son muy diferentes del significado, en el sentido ordinario del término. Apelar a las ideas ordinarias acerca del significado para argumentar a favor de la exclusión del significado estrecho del dominio de la semántica es como excluir a los electrones del dominio de estudio de la materia, basados en que no son "sólidos" y difractan como la luz.

Más aún, el rol del significado estrecho para determinar la función [que va] de contexto a referencia y a valor de verdad parece merecer especialmente el apelativo 'semántica'. (Al discutir más abajo el *Desideratum* 1 argumentaré que el significado estrecho —como lo especifica la semántica de rol conceptual— determina realmente esa función.)

Continuaré hablando, como he hecho, del significado estrecho y del contenido estrecho; pero no me preocupa que el lector prefiera reformular [esa manera de hablar] usando frases como 'determinante estrecho del significado' ['*narrow determinant of meaning*'].

La semántica de rol conceptual y la teoría de dos factores

La semántica de rol conceptual no se encuentra entre los enfoques más populares, pero tiene la peculiaridad de ser el único enfoque (al menos, hasta donde llega mi conocimiento) que posee capacidad [*potencial*], para satisfacer todos los *desiderata*. El enfoque que tengo en mente ha sido sugerido, de modo independiente, tanto por filósofos como por científicos cognitivos: por los primeros, bajo el título de "semántica de rol conceptual", y por los últimos, bajo el título de "semántica procedimental" ["*procedural semantics*"]. (Extrañamente, los dos grupos no se reconocen entre sí.) La doctrina hunde sus raíces en el positivismo, en

el pragmatismo y en la idea wittgeinsteniana del significado como uso. Entre los filósofos, su reciente renacimiento se debe principalmente a Harman (siguiendo a Sellars),[21] y a Field.[22] Churchland, Loar, McGinn y Schiffer han defendido también versiones de este punto de vista.[23] En la ciencia cognitiva, el proponente [*proponent*] principal ha sido Woods,[24] aunque las versiones de Miller y Johnson-Laird[25] han poseído también interés. La versión que a mí me gusta es una "teoría de dos factores" ["*two-factor theory*"] algo así como la teoría defendida por Field,[26] McGinn (1982) y Loar (1982). (Véase también Lycan, 1981.)

La idea de una versión de dos factores es que hay dos componentes del significado, un componente del rol conceptual que está enteramente "en la cabeza" (el significado estrecho)[27] y un componente externo que tiene que ver con la relación entre las representaciones en la cabeza (con sus roles conceptuales internos) y los referentes y/o condiciones de verdad de esas representaciones en el mundo. El enfoque de dos factores deriva del argumento de Putnam (1975, 1979) de que el significado no puede estar a la vez "en la cabeza" y determinar además la referencia. También se inspira en las observaciones de Perry-Kaplan acerca de los deícticos, mencionadas anteriormente (carácter y contenido son los dos "factores"). El enfoque de dos factores puede ser visto como formulando una afirmación conjuntiva [*making a conjunctive claim*] para cada oración: qué es su rol conceptual y cuáles son sus (digamos) con-

21. Véase Harman (1974, 1975 y 1982) y Sellars (1963, 1969 y 1974); véase también Putnam (1979).

22. Véase Field (1977, 1978).

23. Véase Churchland (1979), Loar (1981, 1982), Lycan (1981), McGinn (1982), y Schiffer (1981). Loar y Schiffer defendieron la semántica de rol conceptual sólo como una teoría semántica subsidiaria para el lenguaje del pensamiento, si resulta haber tal cosa. La teoría semántica que defendieron para el lenguaje externo es una teoría griceana funcionalizada.

24. Woods (1977, 1978 y 1981).

25. Johnson-Laird (1977) y Miller y Johnson-Laird (1976).

26. Aunque en un trabajo leído en la Conferencia Sloan del MIT, 1984, Field sugiere un punto de vista según el cual el significado y el contenido son abandonados. Los trabajos de Field de 1977 y 1978 son bastante escépticos respecto de las comparaciones intersubjetivas de rol conceptual, a causa del problema de la información colateral. Por esta razón, dio mucho peso al componente referencial; el escepticismo reciente acerca del componente referencial lo ha conducido al escepticismo respecto del significado y el contenido.

27. Esto es, el aspecto o determinante estrecho del significado.

diciones de verdad.[28] Referiré a la *versión de dos factores* de la semántica de rol conceptual con 'SRC', aunque tal vez debería ser 'SRCDF', para recordar al lector el factor doble que caracteriza a la teoría.

Para los propósitos presentes, la naturaleza exacta del factor externo no es relevante. Quienes se preocupan por ella podrían suponer que es elucidable a través de una teoría causal de la referencia o de una teoría de condiciones de verdad. El factor interno, el rol conceptual, es algo [que concierne al] rol causal de la expresión en el razonamiento y la deliberación y, en general, a la manera en que la expresión se combina e interactúa con otras expresiones para mediar entre los *inputs* sensoriales y los *outputs* conductuales. Un componente crucial del rol conceptual de una oración es la manera [*is a matter*] como participa de las inferencias inductivas y deductivas. El rol conceptual de una palabra es la manera [concerniente a] su contribución en el rol de las oraciones.[29]

Por ejemplo, considérese qué estaría involucrado en que un símbolo en el sistema representacional interno, '→ ', represente al condicional material. El '→, en 'FÉLIX ES UN GATO → FÉLIX ES UN ANIMAL'[30] expresa el condicional material si, por ejemplo, cuando tal oración interactúa apropiadamente con:

> 'FÉLIX ES UN GATO', el resultado es una tendencia a inscribir [*inscribe*] 'FÉLIX ES UN ANIMAL' (en igualdad de circunstancias [*other things equal*], por supuesto).
>
> 'FÉLIX NO ES UN ANIMAL', el resultado es una tendencia a inscribir 'FÉLIX NO ES UN GATO'.
>
> '¿ES FÊLIX UN ANIMAL?' el resultado es una tendencia a iniciar la búsqueda de 'FÉLIX ES UN GATO'.

28. McGinn (1982) estipula la teoría asignando estados de cosas [*states of affairs*] a las oraciones. Esto lleva a LePore y Loewer (1985) a suponer que la teoría de dos factores debe ser más liberal que la teoría de la verdad davidsoniana al permitir, en el factor externo: 'El agua es húmeda' es verdadera ↔ H_2O es húmeda. Pero un teórico de los dos factores *puede* adoptar la teoría davidsoniana de la verdad para el factor externo, aun cuando exigir que la oración del lado derecho del bicondicional sea una *traducción* de la oración citada en el lado izquierdo, constituya un pedido más fuerte de lo necesario para el teórico de los dos factores.

29. Para los propósitos de esta discusión, ignoraré las representaciones internas de carácter pictórico.

30. La escritura cerebral [*brain-writing*] como todo el mundo sabe, se deletrea con letras mayúsculas.

El rol conceptual es el *rol causal total* [*total causal role*] descripto de manera abstracta. Considérese, a modo de analogía, el rol causal del arenque. Los arenques afectan [*affect*] lo que comen, afectan a quienes los comen, a quienes los ven y huyen, y por supuesto, interactúan causalmente unos con otros. Ahora bien, abstraigamos del rol causal total del arenque, su rol culinario; [con 'rol culinario'] quiero significar las relaciones causales que involucran a los arenques, que tienen efecto en la comida humana, o que están afectados por ella. Presumiblemente, algo de lo que afecta a los arenques y que es afectado por ellos, no será parte de su rol culinario; por ejemplo, tal vez los arenques diviertan ocasionalmente a los pingüinos, y esa actividad no tiene causas o efectos culinarios. De modo similar, los elementos del lenguaje tienen un rol causal total que incluye, digamos, el efecto del papel de diario [impreso] en lo que la gente envuelve en él. El rol conceptual abstrae todas las relaciones causales excepto aquellas que median inferencias, inductivas o deductivas, la toma de decisiones [*decision making*] y otras similares.

Una pregunta crucial para la SRC (*la* pregunta crucial) es qué cuenta como identidad y como diferencia de rol conceptual. Claramente, hay muchas diferencias en el razonamiento que no queremos considerar como relevantes para el significado. Por ejemplo, si usted toma más tiempo que yo en razonar de *x* a *y*, no necesariamente querríamos ver esto como revelando una diferencia entre sus significados de *x* y/o *y* y los míos. Nuestros procesos de razonamiento pueden ser los mismos en todos los respectos inferencialmente importantes.

Más aún, la SRC tiene que enfrentar el problema familiar de la "información colateral" [*"collateral information"*]. Supongamos que usted está dispuesto a producir una inferencia de 'TIGRE' a 'PELIGROSO', pero que yo no lo estoy. ¿Tienen nuestros 'TIGREs' el mismo rol conceptual, o no? Más significativamente, ¿qué pasa si diferimos en la inferencia de 'TIGRE' a 'ANIMAL'? ¿Hay una diferencia de tipo entre la primera diferencia y la segunda?

La SRC tiene menos margen de maniobra del que tiene, digamos, la semántica de Katz, dado que la SRC no puede hacer uso de la distinción analítico/sintético. El problema es que si a las inferencias que definen 'gato' las hacemos putativamente analíticas (excluyendo, por ejemplo, la inferencia de 'gato' a 'probablemente capaz de ronronear'), obtenemos un significado de 'gato' que es el mismo que el de 'perro'. (Uno puede tratar de distinguirlos apelando a la diferencia entre las palabras mismas [por ejemplo, el hecho de que 'es un gato' implica 'no es un perro'], pero esto permitiría como mucho la sinonimia intraper-

sonal, no la sinonimia interpersonal. Véase Field, 1978.) Éste no es un problema *dentro* de la semántica de Katz porque los katzianos apelan a elementos primitivos (indefinidos) del lenguaje, en términos de los cuales se definen los demás elementos. (Véase Katz, 1972). El planteo katziano es que uno puede distinguir el significado de 'perro' del de 'gato' apelando a las verdades analíticas de que los gatos son felinos (y no caninos) y los perros son caninos (y no felinos), en donde 'felino' y 'canino' son términos primitivos. Esta movida no es posible para la SRC, dado que no tiene que ver con términos primitivos: se supone que *el rol conceptual determina completamente* al contenido estrecho. (Una salvedad: *es* posible tomar al rol conceptual como una *parte* de la teoría del significado estrecho de *parte* del lenguaje —la parte no primitiva— apelando, al mismo tiempo, a alguna otra concepción del significado de los primitivos; los semánticos procedimentales suenan a veces como si quisieran considerar los términos *fenoménicos* [*phenomenal*] como primitivos, cuyo significado está dado por su "contenido sensorial" ["*sensory content*"], mientras que consideran que otros términos obtienen sus significados *via* sus relaciones computacionales mutuas y también [*via* sus relaciones] con los términos fenoménicos [tal vez ven a los términos fenoménicos como "fundando" las estructuras funcionales]. Debería resultar claro que ésta es un teoría de rol conceptual/fenomenalista [*phenomenalist*] "impura" ["*mixed*"] y no una teoría pura del rol conceptual.

Sin la distinción analítico/sintético, tendríamos que pasar, como mencioné antes, a una concepción científica del significado que suprima la cruda dicotomía mismo/diferente significado, a favor de una gradiente multidimensional de similaridad del significado (esperando resultados tan buenos como aquellos logrados por la teoría de la decisión al pasar de la noción todo-o-nada de la creencia a una noción graduada).

Si la SRC ha de desarrollarse hasta llegar al punto en el que pueda ser evaluada seriamente, tienen que diseñarse e investigarse propuestas definidas para la individuación de roles conceptuales. Uno de los propósitos de este trabajo es tratar de hacer plausible [la tesis de] que la SRC es digna de ser desarrollada.

¿Qué decir de la dimensión social del significado demostrada por Burge (1979)? La teoría de dos factores *puede* tratar de ubicar tal fenómeno en el factor referencial [*referential factor*]. Por ejemplo, tal vez la cadena causal que determina la referencia de mi uso de 'artritis' esté mediada por las actividades de la gente que sabe más de la artritis que yo. (Véase Boyd [1979] para una indicación de cómo tejer el aspecto

social del significado conjuntamente con una teoría causal de la referencia.) Alternativamente, una teoría de dos factores puede expandirse a una teoría de tres factores, dando lugar a un factor social distinto del significado. Dado que mi misión es comparar los rasgos más generales del punto de vista que estoy esbozando con puntos de vista alternativos, no seguiré adelante con el asunto (aunque luego retomaré la pregunta de cómo el factor de rol conceptual [*conceptual role factor*] está relacionado con el factor referencial).

Debería resultar evidente que la SRC, tal como la concibo, está tan poco desarrollada que más que una teoría es sólo el marco para una teoría. ¿Por qué preocuparse en pensar acerca de una teoría tan esquemática? Pienso que el *status* actual de la SRC es parecido al de la "teoría causal de la referencia". La idea básica de las teorías causales de la referencia parece tener una relevancia clara para los fenómenos centrales de la referencia, por ejemplo, cómo una persona puede adquirir de otra persona la aptitud de referir a Napoleón, aunque sin adquirir muchas creencias acerca de Napoleón, y aun cuando gran parte de lo que crea sea falso. Las versiones detalladas de las teorías causales (Devitt, 1981) no han logrado concitar mucho acuerdo. Sin embargo, dado que las únicas teorías disponibles de la referencia (esto es, la teoría de la descripción) no parecen ser defendibles como explicaciones de los fenómenos mencionados, estamos justificados en suponer que las ideas básicas de la teoría causal de la referencia jugarán, de alguna manera, un papel en alguna teoría exitosa de la referencia. Pretendo que los *desiderata* que he discutido proporcionen una razón similar para suponer que las ideas centrales de la SRC tienen que encajar, de algún modo, en nuestro cuadro semántico global.

Debo mencionar que (tal como ocurre con la teoría causal de la referencia) [hay] una semántica de rol conceptual de dos factores [que] ha recibido una versión precisa: la de Field (1977). Aunque la versión de Field es muy sugerente no la adoptaré aquí, por diversas razones. En realidad, el planteo de Field no es una explicación del rol conceptual tal como lo he definido, dado que sus roles conceptuales no son muy causales. Field define el rol conceptual en términos de probabilidad condicional [*conditional probability*]. Dos oraciones tienen el mismo rol conceptual si y sólo si tienen la misma probabilidad condicional respecto de toda otra oración. Aunque Field no lo hace explícito, él se propone [lograr] obviamente, algún tipo de explicación causal en términos de las consecuencias causales de la nueva evidencia en los grados de la creencia...

Aun ignorando este asunto, la explicación de Field llama la atención sobre una elección que los teóricos de la SRC tienen que hacer, una elección que no ha sido sometida a discusión (hasta donde sé), a saber, ¿debe el rol conceptual entenderse en términos ideales o normativos, o debe estar ligado a lo que la gente efectivamente hace? Como Harman señala (en otro contexto), las explicaciones del razonar [*reasoning*] que involucran un cambio del grado de creencia mediante la condicionalización de la evidencia, requieren seguir la pista de un número astronómico de probabilidades condicionales. (Harman calcula que se necesitan mil millones para treinta proposiciones de evidencia [*evidence propositions*].) Así, cualquier explicación bayesiana estaría muy alejada del razonamiento efectivo. Sin embargo, si optamos por ir en contra de tal idealización, ¿debemos transitar tan cerca de la práctica actual como para incluir en el rol conceptual estrategias de razonamiento falaz bien conocidas, tal como la falacia de los apostadores [*gamblers' fallacy*]?[31]

Prefiero no comentar este asunto, en parte porque no estoy seguro de lo que quiero decir y en parte porque estoy tratando de quedar fuera de las controversias [que se dan] *dentro* de la semántica de rol conceptual. Los puntos que quiero plantear pueden formularse sobre la base de una versión de la doctrina que contenga muy pocos detalles.

Llamar 'conceptual' o 'inferencial' a los roles causales a los cuales apela la SRC, no debería llevar a suponer, engañosamente, que la descripción teórica de los mismos puede apelar a sus significados; esto frustraría el objetivo de las teorías reduccionistas. El proyecto que consiste en dar una descripción no-semántica (y no-intencional) de esos roles, es ciertamente intimidatorio [*daunting*], pero el lector hará bien en advertir que no es más intimidatorio que los programas de varias teorías filosóficas populares. Por ejemplo, la teoría causal de la referencia, tomada como una propuesta reduccionista (en la versión de Devitt, no en la de Kripke) soporta el mismo tipo de cargo. Y un cargo bastante similar cae sobre el funcionalismo no-representacional "tradicional" (por ejemplo, en las versiones de Lewis o Putnam), en donde los roles causales de los estados de actitudes proposicionales han de ser descriptos en términos no-intencionales y no-semánticos.

Los representacionalistas difieren en la importancia que atribuyen al rol de las expresiones del español en el razonamiento, la deliberación y demás. En una punta del espectro tenemos la tesis de que el español

31. Véase Kahneman, Slovic y Tversky (1982) y las referencias [que figuran] allí, para estudios detallados de tales falacias.

es *el* lenguaje del pensamiento (para los hablantes del español). Cerca de la otra punta, están quienes, influenciados por la psicología cognitiva, han tendido a ver el razonamiento en términos del español como la punta de un iceberg cuya masa central es la computación en un lenguaje interno [*internal language*] común a los hablantes del español y del walburi.[32] En esta última tesis, el significado estrecho de las expresiones en español es derivado de los significados estrechos de las expresiones del lenguaje interno. (La dependencia, sin embargo, se daría de manera inversa para el componente referencial del significado, dado que son las expresiones del español las que están más directamente relacionadas con el mundo). No me ocuparé de esto ni de una cantidad de disputas que pueden darse *dentro* del marco conceptual de la semántica de rol conceptual.

En lo que sigue, encararé de manera muy laxa el tema del rol del español en el pensamiento. Algunas veces consideraré al español como el lenguaje del pensamiento [*language of thought*]. Sin embargo, cuando sea conveniente, supondré que el español se usa sólo para la comunicación y que *todo* pensamiento es [formulado] en un lenguaje, el mentalés, que no se superpone con el español. Cuando esté en esta última posición, supondré que los mecanismos de producción del lenguaje y de la comprensión del lenguaje establecen una *asociación estándar* [*standard association*] entre el español y las expresiones en mentalés. Cuando un hablante formula un mensaje [*message*] usando 'GATO', los mecanismos de producción del lenguaje mapean [*map*] 'GATO' en 'gato'; y cuando un oyente capta 'gato', los mecanismos de comprensión del lenguaje lo mapean en 'GATO'.

Esta noción de asociación-estándar puede ser usada para construir una manera de individuar los roles conceptuales en los cuales las expresiones en español tienen los roles conceptuales de las expresiones en mentalés con las que están asociadas de manera estándar. Supongamos que me dicen que Félix es un gato y se me pregunta acerca del peso de Félix. Respondo "Félix pesa más de 0,01 gramo". Sugiero comenzar con la siguiente ilustración mecánica simple. Cuando escucho "Félix es un

32. Harman (1970) contrasta los enfoques de la comprensión lingüística de la ruptura de códigos [*code-breaking views of language understanding*] con puntos de vista de incorporativos [*incorporation views*]. En los últimos, la comprensión del inglés es la traducción a un lenguaje diferente; mientras que en los primeros el inglés es parte del lenguaje del pensamiento (en realidad, un sistema de estructuras sintácticas con los ítemes del vocabulario del inglés, es parte del lenguaje del pensamiento), de tal modo, no resulta involucrada ninguna traducción.

gato", los mecanismos de comprensión del lenguaje producen "FÉLIX ES UN GATO". Los mecanismos de razonamiento producen "FÉLIX PESA MAS DE 0,01 GRAMO", y los mecanismos de producción del lenguaje resultan en la proferencia de "Félix pesa más de 0,01 gramo". Ahora bien, una oración en español y su asociada [*associate*] estándar interna, tienen por cierto propiedades causales diferentes. Por ejemplo, una es visible o audible (normalmente) sin técnicas neurofisiológicas. Pero podemos individuar sus roles conceptuales de modo de darles los *mismos* roles conceptuales, simplemente, (1) considerando las propiedades causales relevantes de las expresiones del español como las que están mediadas por sus interacciones causales con sus asociadas estándar, y (2) abstrayendo los mecanismos que efectúan la asociación estándar. Entonces, cualquier causa o efecto de 'gato' será considerada, para los propósitos de la individuación de roles conceptuales, como lo mismo que una causa o un efecto de 'GATO'.

Una analogía: considérese un computador en el que los números son ingresados y mostrados en notación decimal ordinaria, pero en el que toda computación se realiza en notación binaria. La manera en que trabaja el computador es que hay mecanismos que transforman el '3 + 4' que uno ingresa con el teclado, en una expresión interna que podemos representar como '+ (11,100)'. Ésta es una traducción, por supuesto, pero podemos hablar de ella sin describirla como tal, describiéndola en términos del mecanismo que computa la función. Los mecanismos de computación interna operan sobre esta expresión, dando lugar a otra expresión, '111', que se transforma por medio de los mecanismos de traducción, en el '7' desplegado en la pantalla. Ahora bien, el proceso por el cual '3 + 4' da lugar a '7' es exactamente el mismo que el proceso por el cual '+ (11,100)' da lugar a '111', excepto por los dos pasos de traducción. Así, si (1) ignoramos las causas y los efectos de los dígitos decimales distintos de los mediados por sus interacciones con dígitos binarios en las entrañas [*innards*] de la máquina y (2) los abstraemos de los pasos de traducción, podemos considerar a las expresiones decimales y binarias correspondientes como poseedoras de los mismos roles computacionales.

De tal modo, uno puede hablar de los roles conceptuales de las expresiones del español, aun cuando adopte la tesis de que la computación interna es enteramente en mentalés. Esto parecerá extraño si su imagen [*picture*] de los ejemplares del español son expresiones inertes en libros polvorientos, comparadas con las propiedades dinámicas de las representaciones internas en las cuales se conduce efectivamente todo

el pensamiento. Recuérdese que estoy llamando la atención respecto de lo que las expresiones en español hacen cuando son vistas u oídas.

Déjeseme formular algunas advertencias para clarificar lo que estoy tratando de hacer con la noción de asociación estándar.

(1) El español es, por supuesto, un objeto social. Al hablar de los roles conceptuales de las expresiones españolas, no intento [formular] una teoría de ese objeto social. El rol conceptual, se recordará, está destinado a capturar el significado estrecho. Por cierto, dado que los roles causales difieren de persona a persona, la SRC se hace cargo del significado estrecho de *idiolectos* [*idiolect narrow meaning*], más que del significado estrecho del lenguaje público.

(2) La existencia de mecanismos que efectúan la asociación estándar es un problema empírico (aunque, como Stich [1983, pág. 80] argumenta, esa idea parece ser parte, aproximadamente, de la psicología de sentido común). Apelo al trabajo empírico hecho sobre el "módulo lingüístico" [*"language module"*]; véase Fodor (1983). Aunque la suposición empírica se vuelva falsa, una teoría de rol conceptual del (significado estrecho del) lenguaje externo podría aun darse (en términos de las interacciones causales entre el lenguaje externo y el interno), pero lo que se perdería sería la plausibilidad de una teoría de rol conceptual en la cual para casi toda expresión externa, uno podría suponer una expresión interna con el mismo significado estrecho. Así, para ubicar unívocamente mis pretensiones empíricas, déjeseme incluir la suposición de un módulo lingüístico bajo la rúbrica de "representacionalismo".

(3) Para que la noción de asociación estándar sea usable para definir roles conceptuales, debe ser caracterizable no-semánticamente y no-intencionalmente. Pero, ¿no zozobra esta idea ante hechos obvios acerca del tortuoso camino que va del pensamiento al lenguaje; por ejemplo, que la gente miente? Mi apelación al módulo lingüístico es que funciona (una vez activado) sin la intervención de ningún estado intencional. Por supuesto, lo usamos de maneras diversas, dado que al usar el lenguaje tenemos muchos propósitos. El módulo lingüístico funciona igual al mentir y al ser veraz; la diferencia ha de encontrarse en el mensaje en mentalés. Tal vez la confusión sería evitada si uno se centrara en el uso del lenguaje, no en la comunicación, en pensar en voz alta o en los soliloquios internos.

(4) La producción del lenguaje tiene que cargar una parte mayor del peso de la caracterización de la asociación estándar, que la percepción del lenguaje, dado que esta última encuentra complicaciones con los deícticos, y similares. Cuando uno escucha "Estoy enfermo", uno no lo

representa de la manera en que representaría el pensamiento propio de primera persona.

(5) A pesar de la convención que he adoptado de escribir el mentalés con mayúsculas en español, nada en la posición de la SRC requiere que una oración hablada tenga el mismo significado que una oración pensada. Uno puede dar sentido a la idea de que al hablar uno usa la palabra española 'caza' para significar lo que uno significa en el pensamiento con la palabra española 'SILLA'. Imagínese mudándose a un lugar donde se habla un dialecto que difiere del suyo en que intercambia el significado de esas dos palabras. Si continúa pensando en su antiguo dialecto pero habla en el nuevo, usted estaría en la situación descripta. Considérense dos escenarios muy diferentes. En uno, la nueva situación nunca efectúa un cambio en su módulo de producción/percepción lingüística. Al comunicarse, usted ajusta concientemente sus palabras, pero al pensar en voz alta, habla como antes. En el otro escenario, el módulo cambia de tal manera que se ajusta a la mutación externa. En el primer caso la asociación estándar será la normal. En el último, 'silla' estará asociada en manera estándar con 'CAZA', y el rol conceptual de 'silla' derivará de los pensamientos-'CAZA' [*'CHASE'-thoughts*] (involucrando el tratar de capturar en lugar de estar sentado). 'Silla' tendrá el mismo rol conceptual que 'CAZA'. Ningún escenario ocasiona problemas al enfoque del rol conceptual del lenguaje externo, que he esbozado. Schiffer y Loar han enfatizado que si hay un lenguaje interno, una oración hablada [*sentence spoken*] no necesita tener el mismo significado que la misma oración pensada [*sentence thought*], pero han concluido que si la hipótesis del lenguaje del pensamiento es verdadera, es razonable presentar dos tipos bastante diferentes de teorías del significado, una para el lenguaje interno, otra para el lenguaje externo. Su preocupación por el lenguaje externo es por el significado en el lenguaje público, mientras que la mía es por el significado estrecho en un idiolecto; no hay pues un conflicto directo. Quiero enfatizar, sin embargo, que una conclusión análoga a la de ellos, para el significado estrecho de un idiolecto, está errada. (Véase Loar, 1981; Schiffer, 1981). Este tema aparecerá nuevamente en la sección siguiente acerca de qué hace significativas a las expresiones significativas.

Un punto final de clarificación: aunque estoy defendiendo la SRC, estoy lejos de ser un verdadero creyente. Mi posición es que la SRC puede ser suficiente (como está indicado por los *desiderata* que satisface) para motivar que la desarrollemos en detalle y busquemos solucionar sus problemas.

Tal vez éste sea el lugar para mencionar por qué deseo defender una versión del funcionalismo, a pesar de mis argumentos en contra del funcionalismo (Block, 1978). Primero, estoy impresionado por las preguntas que esta versión particular del funcionalismo puede (aparentemente) responder. Segundo, estoy tratando de convenir (y pienso que hay cierto valor) en una teoría que sea chauvinista, en el sentido de que no caracteriza el significado o la intencionalidad en general, sino sólo el significado o la intencionalidad *humanas*. Tercero, los argumentos que di a favor de la conclusión de que el funcionalismo es liberal (en el sentido de que adscribe en demasía propiedades mentales, por ejemplo, a grupos de gente organizada apropiadamente) eran más fuertes contra las teorías funcionalistas de los estados mentales *experienciales* [*experiential mental states*]. Ahora estoy dispuesto a considerar a los estados mentales intencionales como una clase natural para la cual una teoría funcionalista puede ser OK, aun cuando no sea aceptable para los estados experienciales. Por cierto que si el dominio de la SRC es una clase natural, entonces tal es el dominio de los fenómenos mentales intencionales.

Irónicamente, esta concesión al funcionalismo puede hacer que mi posición sea más difícil de defender contra los funcionalistas cabales, dado que me puede comprometer con la posibilidad de intencionalidad —aun de estados intencionales con el mismo tipo de contenido intencional que el nuestro— sin experiencia. Tal vez me vea comprometido con la posibilidad de "zombis", cuyas creencias son las mismas que las nuestras (incluyendo creencias acerca de que tienen dolor), pero que no tienen dolores reales (sólo dolores "artificiales" ["*ersatz*"], que son funcionalmente como los dolores pero que carecen de contenido cualitativo). Tendría, entonces, que afrontar los argumentos en contra de esta posibilidad, dados por Shoemaker (1984, caps. 9 y 14). (Desde mi punto de vista, el dolor, por ejemplo, es efectivamente un estado compuesto consistente en un estado cualitativo no-funcional junto con un estado funcional. Dado que el estado cualitativo puede ser caracterizado neurofisiológicamente, pero no funcionalmente, veo la explicación completa de lo mental como funcional en parte y como fisiológica, en parte.) Finalmente, creo que muchos de los argumentos que se han dado en contra del funcionalismo, en sus varias formas, son defectuosos (véase más abajo mi argumento contra Searle).

¿Dos factores o un factor?

La versión de la SRC de la que he estado hablando es una versión de "dos factores", en la cual el factor rol conceptual está destinado a capturar el aspecto (o determinante) del significado "en la cabeza", mientras que el otro está destinado a capturar las dimensiones referenciales y sociales del significado.

Como mencioné antes, Gilbert Harman ha defendido una versión diferente de la semántica de rol conceptual. La versión de Harman se las arregla [*makes do*] con *un* factor, a saber, el rol conceptual. ¿Cómo lidia [Harman] con los factores referenciales y sociales? Haciendo que su factor único se extienda al mundo de los referentes y de las prácticas de la comunidad lingüística. He estado hablando de los roles conceptuales de acuerdo con las líneas comunes trazadas en los escritos funcionalistas de la filosofía de la mente. Esos roles conceptuales terminan aproximadamente en la piel. Los *outputs* son concebidos en términos de movimientos corporales o, de acuerdo con una mentalidad más científica, en términos de *outputs* de, digamos, el cortex motor (admitiendo [la posibilidad de] pensamientos en cerebros sin cuerpo [*disembodied*]). Los *inputs* se conciben en términos de los estímulos próximos [*proximal*] o en términos de *outputs* de los transductores sensoriales [*sensory transducers*]. En contraste, Harman dice esto sobre el tema.

> La semántica de rol conceptual no involucra una teoría "solipsista" del contenido de los pensamientos. No sugiere que el contenido dependa sólo de las relaciones funcionales entre pensamientos y conceptos, tales como el rol que un concepto particular juega en una inferencia. (Field, 1977, pasa por alto este punto.) También son relevantes las relaciones funcionales con el mundo externo en conexión con la percepción, por un lado, y la acción, por el otro. Lo que hace de algo el concepto de rojo es, en parte, la manera en la que el *concepto* está involucrado en la percepción de los objetos rojos del mundo exterior. Lo que hace de algo el concepto de peligro es, en parte, la manera en la que el concepto está involucrado en los pensamientos que afectan de ciertas maneras la acción.[33]

Uno podría hablar de los roles conceptuales de Harman como "de brazos largos", como opuestos a los roles conceptuales "de brazos cortos" del teórico de los dos factores.

Mi objeción a Harman, en síntesis, es que no veo cómo puede tratar

33. Harman (1982), pág. 14.

los fenómenos que uno pensaría ordinariamente que están al alcance de la teoría de la referencia, sin extender su explicación hasta un punto en el que resulte equivalente a la explicación en términos de dos factores.

El problema surge cuando uno ve las respuestas de Harman a los problemas propios de las teorías familiares de la referencia. Considérese a un residente de la Tierra que viaja a la Tierra Gemela en una nave espacial. Aterriza en un cuerpo de XYZ pero, ignorante de la diferencia entre la Tierra Gemela y la Tierra, emite a su casa el mensaje "Rodeado por agua". A primera vista, uno podría pensar que el rol conceptual harmaniano de la palabra del viajero 'agua' involucraría en ese momento una conexión con XYZ, dado que eso es con lo que su percepción y su acción están conectadas en ese momento. Harman se vería comprometido a decir, entonces, que el mensaje del viajero es verdadero, en contraste con la pretensión de Putnam de que su mensaje es falso porque no está rodeado por agua (sino por agua gemela). Dado que Harman acepta la línea de Putnam, apela a la noción de "contexto normal" ["*normal context*"] (Harman, 1973), cuya idea es que el rol conceptual de 'agua' para el viajero es pensado como involucrando la sustancia a la que él normalmente refiere cuando usa esa palabra.

Otro caso que Harman discute es el caso de Putnam del olmo/haya. (Recordarán que el problema es cómo puedo usar 'olmo' para referirme a los olmos cuando lo que yo sé sobre los olmos es exactamente lo mismo que sé sobre las hayas [excepto por los nombres]). La solución de Harman consiste en incluir en *mi* rol conceptual para 'olmo' su rol en la mente de los expertos que efectivamente conocen la diferencia.

Comienza a parecer como si Harman estuviera construyendo dentro de sus roles conceptuales de brazo largo mecanismos que han sido ubicados usualmente en la teoría de la referencia. El punto puede ser reforzado echando un vistazo a otros fenómenos que han concernido a las teorías de la referencia, tales como tomar prestada [*borrowed*] la referencia de cosas que no existen ahora pero han existido en el pasado. Puedo referirme a Aristóteles sobre la base de haber escuchado por casualidad una conversación acerca de él, aun cuando la mayor parte de lo que creo acerca de Aristóteles sea falso, porque malinterpreté lo que usted dijo. ¿Tratará esto Harman haciendo que sus roles conceptuales se extiendan de una persona a otra en el pasado, esto es, haciendo que una relación causal entre Aristóteles y yo —mediada por usted, y por su palabra, y por la fuente de su fuente [de su palabra], etcétera— sea parte del rol conceptual de mi uso de 'Aristóteles'? Si no, ¿cómo puede tratar Harman la referencia prestada? Si esto es así, Harman nos debe

ciertamente una razón para pensar que la parte exterior-al-cuerpo de sus roles conceptuales de brazo largo, difiere del factor referencial de la teoría de dos factores.[34] El peso de la prueba es para Harman especialmente apremiante, dado que parece que uno podría transformar fácilmente una teoría del tipo de la que él defiende en una teoría del tipo de la que yo defiendo. Si uno toma los roles conceptuales de brazo largo de Harman y "troncha" la porción de estos roles exteriores a nivel de la piel, se queda con mis roles conceptuales de brazo corto. Si la parte exterior-al-cuerpo que es tronchada equivale a algún tipo familiar de teoría de la referencia, entonces la diferencia entre la teoría de un factor de Harman y la teoría de los dos factores es meramente verbal.

La semántica de rol conceptual es tratada a menudo con mofa a causa de no apreciarse la opción que brinda la versión de los dos factores; una falla que es tan común entre los que proponen este punto de vista como entre sus oponentes. Considérese la crítica de Fodor (1978) a la versión de la semántica de rol conceptual de Johnson-Laird. La versión de Johnson-Laird tendía, en su artículo original, al verificacionismo; esto es, los roles de las palabras en los que se centró, eran sus roles en una clase específica de razonamiento, a saber, la verificación. Fodor critica correctamente ese verificacionismo.[35] Pero deseo centrarme en un asunto diferente. Fodor objetó que el significado de 'Napoleón ganó la batalla de Waterloo' no podría consistir posiblemente en ningún conjunto de procedimientos para manipular símbolos internos. Esta idea, argumentó, encarna la falacia uso/mención.

> Supongamos que alguien dice: '¡Qué logro! [*Breakthrough!*] La interpretación semántica de "¿Ganó Napoleón en Waterloo?" es: *Encuentre si la oración "¿Ganó Napoleón en Waterloo?" aparece en el volumen con el número decimal Dewey XXX, XXX en la 42° rama de la Biblioteca Pública de la Ciudad de Nueva York....* " 'Pero', dijo riendo la abuela, 'si eso fuera lo que "¿Ganó Napoleón en Waterloo?" significa, no sería una pregunta acerca de *Napoleón*'. 'Oh, ¡cáspita!', replicó Tom Swift."[36]

34. Véase Loar (1982), págs. 278-80, para un enfoque diferente de lo que está errado en la visión de Harman. Loar toma la línea de que dispositivos tales como los "contextos normales" de Harman y el rol conceptual en las mentes de los expertos, son ad hoc.

35. La respuesta de Johnson-Laird (1978) a Fodor abandona bastante esta tendencia verificacionista a favor de un rol conceptual generalizado similar a la idea a la que he estado aludiendo aquí.

36. Fodor (1978); incluido en Fodor (1981), pág. 211.

La objeción de Fodor es que si el significado se identifica con las interacciones causales de los elementos del lenguaje, las oraciones serían acerca del *lenguaje*, no del mundo.

Mi defensa de Johnson-Laird debería ser obvia. Tómense los procedimientos para manipular 'Napoleón', etcétera (o, mejor, el rol conceptual completo de esas palabras) como especificando el significado *estrecho*. El argumento de Fodor sería dañoso sólo para una teoría que tomara el rol conceptual para especificar aquello *acerca* de lo cual es el lenguaje. Pero si el rol conceptual especifica sólo el significado estrecho, no la referencia o las condiciones de verdad, entonces las críticas de Fodor yerran el blanco. Si Johnson-Laird adoptara una teoría de los dos factores del tipo de la que he estado defendiendo, podría responder a Fodor señalando que la tarea de decir acerca de qué es el lenguaje debería ser tratada por medio del componente referencial de la teoría, no por el componente del significado-estrecho.

Un punto similar se aplica a la crítica bastante pintoresca de Dretske a las observaciones de Churchland y Churchland (1983).

> Suena como si fuera magia: significar algo multiplicando el sonido y la furia. A menos que ponga crema no obtendrá crema helada no importa cuán rápido dé vuelta la manija o cuán sofisticado sea el "proceso". La crema, en el caso de un sistema cognitivo, es el rol *representacional* de aquellos elementos sobre los cuales se realizan [*are performed*] las computaciones. Y el rol representacional de una estructura es, me permito decir, un asunto de cómo los elementos del sistema están relacionados, no los unos con los otros, sino con la situación externa que "expresan".[37]

Pero la crema, de acuerdo con la teoría de dos factores, es el rol conceptual *juntamente con* el rol representacional de Dretske. Dado que la SRC pone la crema de Dretske, con algo *más*, no hay misterio respecto de cómo obtener con ella el helado.

El mismo tipo de observación se aplica a las críticas de la SRC que censuran el componente rol conceptual por no proveer una teoría del significado *completa*. Nuestros juicios acerca de la mismidad [*sameness*] del contenido están controlados por una mezcla compleja de consideraciones acerca del rol conceptual y el referencial (y tal vez por otros).[38]

37. Dretske (1983), pág. 88.

38. Esto está bien argumentado por Stich (1983). (Aunque, como pienso que muestra Sterelny (1985), Stich despliega la noción errada de "potencial" al caracterizar sus roles funcionales.) Extrañamente, Stich considera a las representaciones mentales como indi-

Fodor (1985) señala que el concepto de agua puede ser compartido por mí y mi Yo-Ciego [*Blind Me*]. Dice que esto presenta problemas para teorías como la SRC. Y continúa diciendo:

> La respuesta obvia es que las propiedades de las relaciones causales que hacen a la mismidad y diferencia de roles funcionales son ciertamente muy abstractas. Bueno, puede ser; pero hay una respuesta alternativa que parece mucho menos forzada. A saber que si mi Yo-Ciego puede compartir mi concepto de agua no es porque ambos tengamos representaciones mentales con roles causales abstractamente idénticos; es, más bien, porque ambos tenemos representaciones mentales que están conectadas apropiadamente (digamos, causalmente) con el *agua*.[39]

Pero las dos respuestas que da no son *alternativas incompatibles*; la SRC puede adoptarlas a ambas, aunque pienso que Fodor está en lo correcto respecto de que el hecho de que la referencia efectuada por mí y por mi Yo-Ciego al mismo material, probablemente sea aquí lo principal. El punto es que uno no puede criticar la teoría de dos factores por no hacer todo con un sólo factor.

Panorama

El resto del artículo está dedicado principalmente a mostrar cómo la SRC satisface los *desiderata* y a comparar la SRC con otras teorías semánticas en ese respecto. Hablaré de dos clases muy diferentes (pero compatibles) de teorías semánticas: las reduccionistas y las no-reduccionistas. Una teoría semántica reduccionista caracteriza los términos semánticos en términos no-semánticos. Una teoría semántica no-reduccionista no es la que es *anti*-reduccionista, sino la que no tiene propósitos reduccionistas. Estas teorías están dedicadas principalmente a temas relativos a las construcciones en lenguajes particulares, por ejem-

---

viduadas funcionalmente, sin considerar nunca si hay que hacer alguna distinción entre el aspecto del rol funcional relevante a la semántica y el aspecto que podría ser llamado sintáctico. (Ciertamente, ellos están identificados en la pág. 200.) Ésta es una distinción que hacemos con respecto a la ortografía inglesa. Si alguien escribe la letra 'a' de una manera idiosincrásica, podemos identificarla *funcionalmente*, por la manera en que aparece en las palabras; por ejemplo, aparece sola, aparece en *b*n*n** en el lugar de los asteriscos, etcétera. Al mismo tiempo, podemos distinguir funcionalmente entre dos usos del mismo tipo sintáctico, '*bank*'.

39. Fodor (1985).

plo, por qué 'La temperatura se está elevando' y 'La temperatura es de 70°' no implica '70° se está elevando'. Las teorías no-reduccionistas que mencionaré son las semánticas de mundos posibles [*possible-worlds semantics*], el aspecto teórico modelístico [*model-theoretic*] de la semántica situacional [*situation semantics*], la semántica davidsoniana y la semántica de Katz. Las teorías reduccionistas son la SRC; las teorías griceanas, es decir, teorías que explican la semántica en términos de lo mental, y las que llamo teorías de "indicador" [*"indicator" theories*], cuya metáfora para la semántica es la relación entre un termómetro y la temperatura que indica, o la relación entre el número de anillos del tronco [*stump*] y la edad del árbol cuando se lo cortó. Estas teorías consideran a la relación nomológica entre el indicador y lo que indica como una relación semántica primaria [*prime semantic relation*]. En este campo, incluyo los puntos de vista de Dretske, Stampe, Fodor, y una parte de la posición de Barwise y Perry.

La distinción reduccionista/no-reduccionista, tal como la he bosquejado, no hace justicia a la visión de Davidson. El problema *no* es que la contribución de Davidson acerca de, por ejemplo, la forma lógica de las oraciones de acción lo haga un no-reduccionista, mientras que su enfoque acerca del significado lo hace un reduccionista. Como señalé antes, la empresa reduccionista y la no-reduccionista son compatibles, y nada hay de extraño en que una persona contribuya a ambas. El problema, más bien, es que Davidson tiene un punto de vista acerca de lo que es el significado, que lo hace parecer (erróneamente) como un reduccionista; sin embargo, su enfoque acerca de lo que el significado es, claramente, es *no* reduccionista. (Véase Davidson [1984, pág. xiv], donde describe su proyecto como explicando el significado en términos de la verdad.) Una clasificación más fina distinguiría entre (1a) las teorías reduccionistas y (1b) las teorías no-reduccionistas acerca de lo que el significado es, y distinguiría ambos tipos de enfoques acerca de lo que el significado es, de (2) el proyecto de la semántica teórico modelística, el trabajo de Davidson acerca de las oraciones de acción, y similares. Al etiquetar (1a) como reduccionista, y todo lo demás como no-reduccionista, he amontonado desafortunadamente (1b) y (2), pero esto no es importante para mis propósitos, dado que estoy ignorando las teorías (1b).

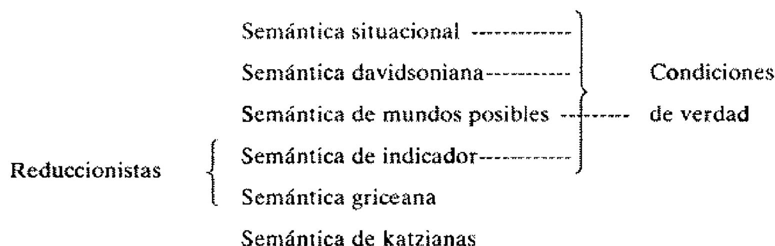
Aunque reduccionista en sus intenciones, la SRC no debe ser vista realmente como compitiendo con las teorías no-reduccionistas. Sin embargo, compararé la SRC con las teorías no-reduccionistas en lo que respecta a los *desiderata* que he registrado. Para prevenir malos enten-

didos, deseo enfatizar que no estoy tratando de criticar las teorías no-reduccionistas. Más bien mi propósito es poner en claro que no deben verse como persiguiendo los mismos objetivos que las teorías reduccionistas.

Compararé la SRC con las teorías reduccionistas. Esas teorías están en el mismo campo de juego que la SRC, pero muchas no son competidoras genuinas. Dado que la SRC, en la versión que estoy promoviendo, es una teoría de dos factores, requiere la compañía de una teoría reduccionista de las condiciones de verdad. La semántica de indicador es un candidato. Otro candidato, que es a la vez de condiciones de verdad y reduccionista, es la interpretación que Field (1972) hace de Tarski. No la discutiré porque no sé que se la haya defendido como una teoría completa del significado; en realidad, Field la ve como un candidato para el factor veritativo-funcional de una teoría de dos factores (véase Field, 1977). Aunque no considero a la semántica de indicador como un competidor real, mencionaré problemas serios relacionados con ese enfoque.

La única circunstancia en la cual las teorías reduccionistas de las condiciones de verdad serían competidoras genuinas de la SRC sería aquella en la cual una de ellas pudiera satisfacer un rango de *desiderata* del tipo que he mencionado. No es problema el que puedan contribuir a ***alguno*** de los ***desiderata***, dado que a menudo hay más de una manera de explicar algo. Pero si alguna teoría de condiciones de verdad los satisficiera *a todos ellos*, la necesidad del componente rol conceptual sería puesto en duda.

[He aquí] una guía breve de las teorías semánticas que mencionaré: reúno las teorías de condiciones de verdad menos la semántica de indicador más la semántica de Katz como no-reduccionistas. Las teorías griceana y de indicador, en contraste, son reduccionistas.



Como se puede ver, cuatro de las seis teorías que compararé con la SRC son clasificables como de condiciones de verdad. Aunque la SRC, en la versión que estoy adoptando, tiene un componente de condiciones de verdad, desempeñará un rol pequeño en la satisfacción de los *desiderata*. Así, puede parecer que estoy censurando a las teorías de condiciones de verdad por no hacer algo que nunca intentaron hacer. La razón para el contraste que trazaré es que el desacuerdo radical es tan común en lo que respecta a asuntos de semántica que hay poco consenso acerca de los propósitos de las teorías semánticas. Para cada una de las teorías de condiciones de verdad que mencionaré, se han hecho alegatos en el sentido de que satisfacen *desiderata* del tipo de los que he registrado.

Representacionalismo

Antes de adentrarme en la discusión de cómo la SRC satisface los *desiderata*, deseo asegurarme de que mi representacionalismo no sea mal comprendido. Estoy comprometido con que el razonamiento complejo [*complex reasoning*] es un proceso que involucra la manipulación de estructuras simbólicas. No estoy comprometido con la idea de que esas estructuras simbólicas sean *independientes* de los estados representacionales de la mente, de que los objetos mentales sean vistos por un ojo interior. Es conveniente hablar en términos de representaciones internas como si fueran literalmente oraciones en el cerebro (y yo hablo de esa manera), pero este hablar es, por supuesto, metafórico. Mi compromiso será satisfecho si los estados representacionales mismos constituyen un sistema combinatorio; esto es, si son estructurados de tal manera que las partes correspondientes a las palabras se pueden combinar de modo de constituir estados representacionales que correspondan a oraciones.[40]

No estoy comprometido con que la manipulación de estructuras de símbolos esté involucrada en *todo* razonamiento, dado que quiero permitir [la existencia de] razonamientos primitivos a partir de los cuales se construyen los razonamientos complejos. (Por ejemplo, en algunos computadores, la multiplicación es un proceso simbólico en el que un problema de multiplicación se "descompone" en series de problemas de

40. Véase Hills (1981), págs. 18-19, para una discusión de dos maneras de hablar del simbolismo interno, y Harman (1973) para una aplicación de la versión del estado representacional.

adición, pero la adición misma no se "descompone" en otro tipo de problemas, sino más bien se cumple por medio de un dispositivo del *hardware*, un procesador primitivo, que no contiene representaciones internas. Si uno se pregunta cómo multiplica la computadora, se obtiene una respuesta representacional, si se pregunta cómo suma, no.) No estoy comprometido con reglas de razonamiento que sean ellas mismas representadas. Tal suposición involucra paradojas notables y en los computadores tenemos ejemplos de manipuladores de símbolos muchas de cuyas "reglas" para la manipulación de símbolos están implícitas en la manera en que el *hardware* trabaja (véase Block, 1983). No estoy comprometido con ninguna tesis detallada acerca de cómo son las computaciones internas. Por ejemplo, no estoy comprometido con la idea de que al computar '99 + 99 = 198' haya algún análogo interno a "llevarse un 1", o cualquiera otra manipulación de símbolos del tipo de las que una persona podría realizar al hacer tal suma.

Más aún, la pretensión de que somos manipuladores de símbolos intenta ser empírica y contingente. Encuentro la idea de que somos computadores "analógicos" [*"analog" computers*] cuyas actividades internas no involucran para nada manipulaciones de símbolos, perfectamente inteligible y plausible. Formulo el supuesto representacionalista por dos razones: la línea de investigación más promisoria en la ciencia cognitiva está masivamente comprometida con el representacionalismo, y parece dar buenos resultados, y creo que hay un número astronómico de pensamientos que la gente es capaz de tener. Yo argumentaría que el número de oraciones pensables de treinta palabras de largo es más grande que el número de partículas en el universo. Considérese un conjunto de oraciones susceptibles de ser tomadas en consideración [*entertainable*] que tenga la siguiente forma: $n \times m = q$, donde n y m están en el rango familiar de cientos de miles de millones del presupuesto nacional (doce dígitos), y q es el doble de largo. Muchas de estas oraciones no son creíbles (por ejemplo, novecientos mil millones multiplicado por sí mismo = 0), pero cada una es ciertamente pensable. El número de proposiciones distintas tomadas en consideración, de la forma mencionada, está en el orden de los cuarenta y seis dígitos de largo. Una comparación instructiva: el número de segundos desde que comenzó el tiempo es sólo de aproximadamente dieciocho dígitos de largo. No veo qué mecanismo podría ser aquel por el cual una persona pueda pensar uno cualquiera de tal vasta variedad de pensamientos, sin algún tipo de sistema combinatorio involucrado. Mi supuesto representacionalista está en el espíritu de la pretensión de Smart de que el dolor es un estado

cerebral: una tesis basada empíricamente acerca de cómo parece ser el razonamiento.

De las teorías semánticas que contrastaré con la SRC, sólo la versión de la semántica de indicador de Fodor tiene un supuesto representacionalista comparable; sin embargo, no pienso que mi representacionalismo tenga que ser visto como la diferencia clave entre la teoría que estoy defendiendo y la mayoría de las otras teorías. Por un lado, una teoría denotacional como la de Fodor podría enmarcarse en términos del asentimiento a oraciones del español en lugar de relaciones computacionales con oraciones internas. Fodor es oracionalista [*sententialist*] porque cree que los estados de actitudes proposicionales son relaciones con oraciones internas. Pero las oraciones internas no tienen un rol *semántico* privilegiado en su explicación. Además, no hay caminos no-representacionalistas que conduzcan al tipo de semántica basada en el funcionalismo que estoy defendiendo; por ejemplo, las versiones de Loar y Schiffer del programa griceano. Si la SRC en la forma en la que estoy defendiéndola se encontrara con problemas empíricos serios a causa de su representacionalismo, intentaría una versión no-representacionalista.

Pregunta: si mi compromiso básico es con una teoría funcionalista del significado, ¿por qué no adopto *ahora* una versión no-representacionalista del funcionalismo (por ejemplo, el programa Loar-Schiffer) en lugar de intentar un programa basado en una suposición empírica riesgosa (el representacionalismo)? Respuesta: como señalaré más adelante, aun cuando el programa Loar-Schiffer funcione para los lenguajes naturales, si hubiera un lenguaje del pensamiento que no fuera idéntico al lenguaje natural, su teoría no funcionaría para *él*. Así, *ambas* teorías están sujetas a un riesgo empírico. El suyo es inadecuado si el representacionalismo es verdadero, mientras que el mío está equivocado si el representacionalismo es falso...

TRADUCTORA: Diana Pérez.
REVISIÓN TÉCNICA: Eduardo Rabossi.

REFERENCIAS BIBLIOGRÁFICAS

Block, Ned: (1978) "Troubles with Functionalism", en *Perception and Cognition: Issues in the Foundations of Psychology*, C.W. Savage (comp.), Minneapolis.

Block, Ned: (1983) "Mental Pictures and Cognitive Science", *Philosophical Review* 92, 499-541.
Boyd, Richard: (1979) "Metaphor and Theory Change", en *Metaphor and Thought*, Andrew Ortony (comp.), Cambridge.
Burge, Tyler: (1979) "Individualism and the Mental", *Midwest Studies in Philosophy* 4, 73-121.
Burge, Tyler: (1984) "Individualism and Psychology", *Philosophical Review*.
Churchland, Paul M.: (1979) *Scientific Realism and the Plasticity of Mind*, Cambridge.
Davidson, Donald: (1984) *Truth and Interpretation*, Oxford.
Devitt, Michael: (1981) *Designation*, Nueva York.
Dretske, Fred I.: (1983) "Why Information?", *Behavioral and Brain Sciences* 6, 82-89. Ésta es la respuesta de Dretske a sus críticos.
Field, Hartry: (1972) "Tarski's Theory of Truth", *Journal of Philosophy* 69, 347-75.
Field, Hartry: (1977) "Logic, Meaning, and Conceptual Role", *Journal of Philosophy* 74, 379-409.
Field, Hartry: (1978) "Mental Representation", *Erkentniss* 13, 9-61.
Field, Hartry: (1980) *Science without Numbers: A Defense of Nominalism*, Oxford.
Fodor, J.A: (1978) "Tom Swift and His Procedural Grandmother", *Cognition* 6, 229-247.
Fodor, J.A.: (1981) *Representations*, Cambridge, Mass.
Fodor, J.A.: (1983) *The Modularity of Mind*, Cambridge, Mass.
Fodor, J.A.: (1985) "Banish Discontent", en *Proceedings of the 1984 Thyssen Conference*, Jeremy Butterfield (comp.), Cambridge.
Harman, Gilbert: (1970) "Language Learning", *Noûs* 4, 33-43. Incluido en *Readings in Philosophy of Psychology*, vol. 2, Ned Block (comp.), Cambridge, Mass.
Harman, Gilbert: (1973) *Thought*, Princeton, N.J.
Harman, Gilbert: (1974) "Meaning and Semantics", en *Semantics and Philosophy*, M.K. Munitz y Peter Unger (comps.), Nueva York.
Harman, Gilbert: (1975) "Language, Thought and Communication", en *Language, Mind and Knowledge*, K. Gunderson (comp.).
Harman, Gilbert: (1982) "Conceptual Role Semantics", *Notre Dame Journal of Formal Logic* 23, 242-56.
Haugeland, John: (1980) "Programs, Causal Powers, and Intentionality", *Behavioral and Brain Sciences* 3, 432-33.
Hills, David: (1981) "Mental Representations and Languages of

Thought", en *Readings in Philosophy of Psychology*, vol. 2, Ned Block (comp.), Cambridge, Mass.

Horwich, Paul: (1982) "Three Forms of Realism", *Synthèse* 51, 181-201.

Johnson-Laird, P.N.: (1977) "Procedural Semantics", *Cognition* 5, 189-214.

Johnson-Laird, P.N.: (1978) "What's Wrong with Grandma's Guide to Procedural Semantics: A Reply to Jerry Fodor", *Cognition* 6, 241-61.

Kahneman, D.P. Slovic, y A. Tversky: (1982) *Judgement under Uncertainty: Heuristics and Biases*, Cambridge.

Kaplan, David: (1980) "Demonstratives". Texto mimeografiado, John Locke Lectures.

Katz, Jerrold J.: (1972) *Semantic Theory*, Nueva York.

Kripke, Saul: (1979) "A Puzzle about Belief", en *Meaning and Use*, A. Margalit (comp.), Dordrecht.

LePore, E., y B. Loewer: (1985) "Dual Aspect Semantics".

Loar, Brian: (1981) *Mind and Meaning*, Cambridge.

Loar, Brian: (1982) "Conceptual Role and Truth Conditions", *Notre Dame Journal of Formal Logic* 23, 272-83.

Lycan, W.: (1981) "Toward a Homuncular Theory of Believing", *Cognition and Brain Theory* 4, 139-59.

McGinn, Colin: (1982) "The Structure of Content", en *Thought and Object*, Andrew Woodfield (comp.), Oxford.

Miller, G.A., y P.N. Johnson-Laird: (1976) *Language and Perception*, Cambridge, Mass.

Perry, John: (1977) "Frege on Demonstratives", *Philosophical Review* 86, 474-97.

Perry, J.: (1979) "The Problem of the Essential Indexical", *Noûs* 13, 3-21.

Putnam, Hilary: (1975) "The Meaning of 'Meaning'", en *Language, Mind, and Knowledge*, K. Gunderson (comp.), Minneapolis. También en Putnam, *Mind, Language and Reality*.

Putnam, Hilary: (1979) "Reference and Understanding", en *Meaning and Use*, Avishai Margalit (comp.), Dordrecht.

Pylyshyn, Zenon: (1984) *Computation and Cognition*, Cambridge, Mass.

Schiffer, Stephen: (1981) "Truth and the Theory of Content", en *Meaning and Understanding*, H. Parret (comp.), Berlín.

Searle, John: (1980) "Minds, Brains and Programs", *Behavioral and Brain Sciences* 3, 417-24.

Sellars, Wilfrid: (1963) *Science, Perception and Reality*, Londres. Véase "Empiricism and the Philosophy of Mind" y "Some Reflections on Language Games."

Sellars, Wilfrid: (1969) "Language as Thought and as Communication", *Philosophy and Phenomenological Research* 29, 506-27.

Sellars, Wilfrid: (1974) "Meaning as Functional Classification", *Synthèse* 27, 417-27.

Shoemaker, Sydney: (1984) *Identity, Cause, and Mind*, Cambridge.

Sterelny, Kim: "Is Semantics Necessary? Stephen Stich's Case Against Belief", *Australasian Journal of Philosophy*, 56.

Stich, Stephen: (1983) *The Case Against Belief*, Cambridge, Mass.

Woods, William: (1977) "Meaning and Machines", en *Proceedings of the International Conference on Computational Linguistics*, A. Zampoli (comp.), Florence.

Woods, William: (1978) *Semantics and Quantification in Natural Language Question Answering*, Technical report 3687, Cambridge, Mass.

Woods, William: (1981) "Procedural Semantics as a Theory of Meaning", en *Elements of Discourse Understanding*, A. Joshi, B. Webber, y I. Sag (comps.), Cambridge.

Capítulo 12

UN ARGUMENTO MODAL EN FAVOR DEL CONTENIDO ESTRECHO *

Jerry A. Fodor **

He aquí una antinomia moderna. De un lado, existe el argumento *A*:

Argumento A:

1. Mi gemelo y yo somos duplicados moleculares.
2. Por lo tanto, nuestras conductas (reales y contrafácticas) son idénticas en los aspectos relevantes.
3. Por lo tanto, los poderes causales de nuestros estados mentales son idénticos en los aspectos pertinentes.
4. Por lo tanto, mi gemelo y yo pertenecemos a la misma clase natural a los efectos de la explicación psicológica, y el "individualismo" es verdadero.

Pero, del otro lado, existe el argumento *B*:

Argumento B:

1'. Mi gemelo y yo somos duplicados moleculares.
2'. Sin embargo, nuestras conductas (reales y contrafácticas) son diferentes en los aspectos pertinentes.

* "A modal argument for narrow content", *The Journal of Philosophy*, 88 (1991), págs. 5-25. Con autorización del autor y del *Journal of Philosophy*.

** Estoy profundamente agradecido por las conversaciones mantenidas con George Rey y Steven Stich sobre los temas de este artículo. Reconozco asimismo la ayuda de Ned Block, Anne Jacobson, Tim Maudlin, Colin McGinn, Brian McLaughlin y Stephen Schiffer, así como también los constructivos comentarios de Fred Adams, Joe Levine y David Rosenthal.

3'. Por lo tanto, los poderes causales de nuestros estados mentales son diferentes en los aspectos pertinentes.
4'. Por lo tanto, mi gemelo y yo pertenecemos a clases naturales diferentes, y el "individualismo" es falso.

Al menos uno de estos argumentos tiene que ser erróneo. ¿Cuál lo es? ¿Y qué es lo que en él es erróneo?

En el capítulo 2 de *Psychosemantics*,[1] ofrecí algunas consideraciones a favor del argumento *A*. Se suponía que mostraban que los estados mentales que sólo difieren en las propiedades intencionales "amplias" ["*broad*"] (los tipos de propiedades intencionales que los estados mentales de los gemelos moleculares pueden no compartir) no difieren ipso facto en los poderes causales, y por ello, que las meras diferencias en el contenido intencional amplio no determinan diferencias en las clases naturales a los efectos de la explicación psicológica. Sin embargo, los argumentos que ofrecí en *Psychosemantics* no fueron recibidos con mucho entusiasmo. Era tal su sutileza que yo mismo no siempre estuve seguro de cómo funcionaban. En este artículo, me propongo apoyar otra vez el argumento *A*. Usaré muchos de los materiales que usé en *Psychosemantics*, pero juntaré las piezas de modo diferente. Diré, al pasar, un par de palabras acerca de algunos comentarios provocados por los argumentos de *Psychosemantics*.

No obstante, antes de emprender ese camino formularé una observación preliminar. Para muchos propósitos filosóficos puede no importar demasiado qué resulte del tema del individualismo. Por ejemplo, no pienso que una decisión a favor del argumento *A* (esto es, a favor del individualismo) afecte el *status* del "externalismo" ["*externalism*"] en la semántica. El externalismo es independiente del individualismo porque, cualquiera que sea el *status explicativo* del contenido amplio [*broad content*], no está en discusión (en modo alguno lo *está*) que el contenido de los pensamientos de mi gemelo acerca de *agua* difiera del contenido de los míos; o que "agua" signifique algo diferente en mi boca que en la suya; o que esas diferencias semánticas deriven de diferencias en nuestras respectivas relaciones cabeza/mundo. (Presumiblemente, derivan del hecho de que mientras que la historia causal de mis pensamientos acerca de *agua* los conecta con muestras de H_2O, su historia causal los conecta con muestras de XYZ.) Dicho brevemente, si las intuiciones acerca de los gemelos favorecieran el contenido externalista, entonces

1. Cambridge, MIT, 1987.

queda en pie la cuestión de si los estados intencionales amplios determinan o no a las clases naturales, a los fines de la explicación causal en psicología.

Asimismo —contra lo que sostiene, por ejemplo, Steven Stich[2]— no es obvio, en modo alguno, que la vindicación científica del realismo intencional dependa de cómo se resuelven los temas del individualismo. Supongamos que existiera algún argumento metafísico que mostrase que la explicación causal en psicología requiere que la individuación de los estados mentales sea individualista (de modo que los estados mentales de los gemelos moleculares sean ipso facto idénticos en tipo [*type-identical*], como en el argumento *A*). Se seguiría entonces que la explicación psicológica no es, estrictamente hablando, una especie de explicación en términos de deseos/creencias, puesto que se acepta que los deseos y las creencias se individuan de manera amplia. Pero quedaría pendiente la cuestión de decidir si las explicaciones psicológicas son especies de explicaciones *intencionales*. Eso dependería de si puede construirse una noción individualista de "estado intencional" que sea adecuada a los fines de la explicación psicológica. (Para más información sobre esto, ver *Psychosemantics*.)

En pocas palabras, se pueden hacer muchas cosas en la semántica y en la filosofía de la mente, dejando a un lado los problemas relativos al individualismo. Pienso, sin embargo, que vale la pena tratar de poner esos problemas en orden. Veremos que plantean algunas cuestiones interesantes acerca de la familia de nociones que incluye a la explicación causal, el poder causal, la identidad de tipo y las clases naturales. En consecuencia, deshacer esta madeja puede serle útil a la metafísica y a la filosofía de la ciencia, aun si dejara las cosas más o menos inalteradas en la semántica y en la filosofía de la mente. Comencemos, pues.

El argumento *A* dice que en virtud de nuestra identidad molecular la conducta de mi gemelo y la mía son idénticas *en todos los aspectos pertinentes*. Por supuesto, no es que sean idénticas *tout court*. Por el contrario, del mismo modo que se acepta en esta discusión que los *estados mentales* de los gemelos pueden diferir en algunas de sus propiedades intencionales, también se acepta que las *conductas* de los gemelos puedan diferir bajo alguna de sus descripciones intencionales. Podría ser que la segunda concesión fuera implicada por la primera, puesto que es plausible sostener que las propiedades intencionales de las conductas se heredan de los contenidos intencionales de sus causas mentales. Esto es

2. *From Folk Psychology to Cognitive Science* (Cambridge, MIT, 1983).

plausible porque yo puedo pensar acerca del agua y mi gemelo no puede, yo puedo buscar agua, extraerla, o recomendarle beber agua a un amigo sediento, pero mi gemelo no puede. Un individualista convencido podría argumentar que, *stricto sensu*, nada de esto constituye realmente un ejemplo de descripción *conductual*; pero éste no es el rumbo que me propongo tomar.

Así pues, los gemelos pueden diferir en los contenidos de sus estados mentales; y como consecuencia de esas diferencias en los estados mentales, las conductas de los gemelos pueden diferir en algunas de sus propiedades intencionales. Daré por supuesto (lo cual es, sin embargo, discutible) que esas diferencias en las propiedades intencionales son las *únicas* diferencias (pertinentes) entre las conductas de los gemelos. Es decir, estoy dando por supuesto que las conductas de los gemelos son ipso facto idénticas en todas las propiedades no intencionales que son relevantes para la taxonomía psicológica. Por supuesto que esto es más fuerte que suponer lo que literalmente todo el mundo acepta, a saber, que nuestras conductas son idénticas bajo descripciones *físicas*, esto es, idénticas *qua* movimientos.

He aquí una aparente excepción a la afirmación más fuerte: si profiero [*utter*] "Denme agua" entonces, si todo sale bien, consigo agua; pero si mi gemelo profiere "Denme agua" entonces, si todo sale bien, él consigue "gagua" [*twater*]. Así, puede ser que existan descripciones no intencionales bajo las cuales nuestras conductas sean, después de todo, diferentes de modos pertinentes. Sin embargo, pensándolo bien, esas diferencias desaparecen. Es una muestra de aburrida lucubración que mientras que mi proferencia ocurre en la tierra, la de mi gemelo ocurre en la Tierra Gemela. La lucubración que interesa es que si mi gemelo hubiera proferido "Denme agua" aquí, hubiera conseguido agua, y que si yo hubiera emitido "Denme agua" en la Tierra Gemela, hubiera conseguido gagua. Para plantearlo de modo distinto, mi "Denme agua" no logra agua en cualquier caso; con suerte, logra agua *en ciertas circunstancias*. Pero en *esas* circunstancias, su "Denme agua" también consigue agua.[3] La moraleja es que uno tiene que juzgar la

3. He aquí un caso relacionado —sugerido por un ejemplo de Colin McGinn— que considero que es susceptible del mismo tipo de tratamiento. Supóngase que en la Tierra-Gemela exista no sólo agua-gemela sino también sal-gemela (que es LCaN en vez de CLNa). Y supóngase que el hecho de pensar acerca de la sal provoque su deseo de agua (mientras que pensar acerca de LCaN provoca en su gemelo el deseo de XYZ). De este modo, ¿no habría entonces una diferencia entre los poderes causales de los pensamientos [acerca] de sal y los pensamientos [acerca] de sal del gemelo, en virtud de las diferencias

identidad y las diferencias de los poderes causales de modo de tener presentes los contrafácticos, a través [*across*] de los contextos antes que dentro [*within*] de los contextos.[4] Eso, sin embargo, no nos pone a salvo. Porque adviértase que en el caso en que emito "Denme agua" en la Tierra Gemela y mi gemelo la emite aquí, ninguno de los dos consigue lo que pide. Esto es, sea cual fuere el contexto de la proferencia, mi proferencia es un pedido de agua y su proferencia es un pedido de gagua. Así, nuestras conductas permanecen diferentes de modo pertinente bajo esas descripciones intencionales aun para *el test intercontextual* [*the cross-context test*]. Es esta diferencia residual entre las conductas —su diferencia intercontextual [*cross-context difference*] bajo ciertas descripciones intencionales— lo que constituye el desafío al individualismo y a la superveniencia [*supervenience*] local.

Así pues, la cuestión respecto del individualismo es: ¿pertenecen los estados mentales de los gemelos a clases naturales diferentes (tienen ellos diferentes poderes causales) en virtud de las diferencias en las pro-

---

entre los contenidos de los deseos que ellos causan? No, porque los pensamientos [acerca] de sal carecen del poder para causar deseos [acerca] de agua, en lo que fuere; lo que tienen es el poder de causar deseos [acerca] de agua *en alguien que tiene el concepto de agua.* (Esto es, en alguien que tiene las conexiones causales/históricas apropiadas —o lo que fuere— con H_2O.) Pero, por supuesto, los pensamientos [acerca] de sal de los gemelos tienen también ese poder (del mismo modo que los pensamientos [acerca] de sal tienen el poder de causar deseos [acerca] de gagua en alguien que tenga las conexiones causales/históricas adecuadas con XYZ). Aquí, como en cualquier otro lado, uno aplica el test intercontextual preguntando si *A* tendría los mismos efectos que tendría *B* si *A* interactuase con las mismas cosas (en este caso, con las mismas cosas mentales) con las que *B* ciertamente interactúa.

El resultado es: el test intercontextual muestra que *ciertas diferencias que tienen sus efectos bajo ciertas descripciones intencionales* (a saber, causar deseos de CLNa *versus* causar deseos de LCaN) y *ciertas diferencias que sus efectos tienen bajo una descripción no-intencional* (ver texto) *no* hacen que la diferencia entre tener pensamientos [acerca] de agua y tener pensamientos [acerca] de gagua sea una diferencia de poder causal. Sin embargo, vamos a ver que hay diferencias entre los efectos que exhiben los estados mentales de los gemelos bajo descripciones intencionales, que *sobreviven* al test intercontextual. Éstas son las que, prima facie, plantean problemas al individualismo.

4. Es seguro que ésta es la manera intuitivamente natural de comparar los poderes causales. Considérese:

— "Los gatos criados en Manhattan son incapaces de subir a los árboles", dice un importante científico.

— ¡Caramba! Es extraño. ¿Cómo lo explica?

— No hay ningún árbol en Manhattan adonde se puedan subir.

— ¡Oh!

[...]

piedades intencionales de la conducta de los gemelos de las que son responsables? Dado que "poder causal", "clase natural" y cosas semejantes son, por supuesto, términos técnicos, esta cuestión no se caracteriza por tener un alto grado de claridad. Sin embargo, veremos que existen algunas intuiciones claras acerca de los casos y que ellas nos servirán para nuestros propósitos.

Debería ser evidente que el tipo de cuestión que estamos planteando respecto de los estados intencionales de los gemelos también puede surgir en casos que no tengan nada que ver, en particular, con la intencionalidad. Supongamos que tenemos un par de causas *C*1, *C*2, junto con sus respectivos efectos *E*1, *E*2. Supongamos que:

> *C*1 difiere de *C*2 en que *C*1 tiene la propiedad-causa [*cause property*] *PC*1 en tanto que *C*2 tiene la propiedad-causa *PC*2.
>
> *E*1 difiere de *E*2 en que *E*1 tiene la propiedad-efecto [*effect property*] *PE*1 y *E*2 tiene la propiedad-efecto *PE*2.
>
> La diferencia entre *C*1 y *C*2 es la responsable de la diferencia entre *E*1 y *E*2 en el sentido de que si *C*1 hubiera tenido *PC*2 en vez de *PC*1, entonces *E*1 habría tenido *PE*2 en vez de *PE*1, y si *C*2 hubiera tenido *PC*1 en vez de *PC*2, *E*2 habría tenido *PE*1 en vez de *PE*2.[5]

Denominemos *S* a este esquema.[6] Y ahora lo que queremos saber es: ¿qué instancias [*instances*] del esquema *S* son casos en los que la diferencia entre tener *PC*1 y tener *PC*2 es una diferencia en el poder causal en virtud de su responsabilidad [*responsibility*] en [producir] la diferencia entre *E*1 y *E*2? (a menudo reduciré esto a "¿cuándo tener *PC*1 en vez de *PC*2 es un poder causal?"). Si conociéramos la respuesta a esta pregunta general, entonces sabríamos, en particular, si la diferencia entre tener pensamientos [acerca] de agua [*water thoughts*] y tener pen-

5. A fin de que las cuestiones relevantes no estén presupuestas, es importante que esto sea todo lo que se requiere para que el hecho de que C1 tenga *PC*1 sea "responsable de" el hecho de que *E*1 tenga *PE*1 (y de manera similar, mutatis mutandis, para que el hecho de que C2 tenga *PC*2 sea responsable del hecho de que *E*2 tenga *PE*2). En un sentido similar, nada en los ejemplos dependerá de enfatizar el requisito de que las *C*s y los *E*s estén relacionados como causas y efectos, en tanto se suponga que la diferencia entre las *C*s es responsable de la diferencia entre los *E*s, en el sentido que acabo de especificar.

6. Será fácil la exposición si pensamos al esquema *S* a veces relacionando eventos y a veces relacionando eventos-tipo. Explotaré esta ambigüedad en lo que sigue, pero nada en el argumento depende de esto.

samientos [acerca] de gagua es una diferencia en poder causal, en virtud de ser responsable de la diferencia entre mi producir conductas [referentes a] agua [*water behaviors*] y el producir [por parte] de mi gemelo conductas [referentes a] gagua.[7] Y, si supiéramos *eso* sabríamos si el individualismo es verdadero; que es lo que al comienzo queríamos saber, como el paciente lector recordará.

Ahora bien, una primera reacción plausible a esto es [considerar] que no se trata de un tema filosófico. Porque, podría decirse, la cuestión que se plantea es cuándo el hecho de que una generalización que apoye contrafácticos, muestra que tiene un *status* explicativo. Y la respuesta dependerá, como es habitual, de cuestiones sistemáticas respecto de la simplicidad, la plausibilidad, el poder, etcétera de las explicaciones, y de si disponemos de generalizaciones alternativas que den apoyo a contrafácticos [*alternative counterfactual supporting generalizations*]. Algunas de las cosas que Tyler Burge dice acerca de la conveniencia de no razonar a priori respecto de la taxonomía que la explicación psicológica requiere, sugieren que aprobaría esta línea de pensamiento; por ejemplo:

> Es un error... permitir que preconcepciones ontológicas que sólo tienen apuntalamiento filosófico afecten nuestra interpretación de las empresas científicas. Es un error aun mayor permitirles que dicten los tipos de clases explicativas que se juzgan admisibles para la explicación.[8]

Pero aun cuando no pienso que el apriorismo esté dentro de mis vicios metodológicos mayores, esto me suena un poco rebuscado. Sería sorprendente que pudiéramos ofrecer a priori condiciones *suficientes* interesantes para [determinar] cuándo una diferencia en las propiedades de las causas constituye una diferencia en sus poderes causales. Pero parece probable que se puedan dar argumentos a priori para algunas condiciones *necesarias*. Después de todo, el compromiso con la explicación causal tiene, presumiblemente, *algunas* consecuencias metodológicas *qua* compromiso con la explicación causal, y debe ser posible des-

7. Recordatorio: 'Conducta [referente a] agua' no significa *conducta que tiene que ver con agua*, sino *conducta que hace referencia a agua en una descripción intencional*. Supongo (ver más arriba) que sólo las descripciones relevantes bajo las cuales difieren las conductas de los gemelos, son intencionales.

8. "Individualism and Causation in Psychology" (manuscrito, UCLA, 1989).

lindarlas si reflexionamos acerca de qué tipos de cosas son las explicaciones causales.

Y en realidad parece razonablemente claro, a priori, que algunas de las instancias del esquema *S* no son genuinas [*bona fide*]. Considérese, como un ejemplo totalmente trivial, el caso en el que *PE*1 es la propiedad de ser el efecto de *C*1 y *PE*2 es la propiedad de ser el efecto de *C*2. Existe, por supuesto, una propiedad de *C*1 en virtud de la cual sus efectos son los efectos de *C*1, a saber, la propiedad de ser *C*1. Concordantemente, existe una propiedad de *C*2 en virtud de la cual sus efectos son los efectos de *C*2, a saber, la propiedad de ser *C*2. El apoyo contrafáctico se da de la manera requerida; si la causa de *E*1 hubiera tenido la propiedad de ser *C*2 en vez de *C*1, entonces *E*1 hubiera tenido la propiedad *PE*2 en vez de la propiedad *PE*1. Pero parece obvio a priori que éste no es un caso en el que tener *PC*1 sea un poder causal de las *C*s en virtud de su responsabilidad de que los *E*s tengan *PE*1. Una de las propiedades de mis efectos, que tus efectos no pueden tener por más que lo intentes, es *la propiedad de ser causados por mí*. Pero considero obvio que esta diferencia en nuestros efectos no torna a la propiedad de *ser yo más bien que tú*, un poder causal. No soy una clase natural unitaria en virtud de mi poder singular de causar efectos que sean efectos míos.

Parece claro a priori, entonces, que no todo caso en el cual una diferencia en las causas es responsable de una diferencia en los efectos, sea un caso en el que la diferencia en las causas es una diferencia en sus poderes causales. Existen, en realidad, muchos ejemplos. Puedo definir la propiedad *ser una partícula H* que es satisfecha por cualquier *x* en el tiempo *t* si y sólo si [(*x* es una partícula física en *t*) & (la moneda en mi mano, en *t*, está "de cara")]. Y de manera correspondiente para las partículas T. Así, una diferencia entre las propiedades de las monedas en mi mano (a saber, la diferencia entre estar "de cara" y estar "de cruz") es responsable de la diferencia entre el estado de cosas en el cual todas las partículas en el universo son partículas H y el estado de cosas en el cual todas las partículas en el universo son partículas T. Pero, por supuesto, la diferencia entre estar de cara y estar de cruz no cuenta como un poder causal en virtud de ser responsable de esta diferencia en las partículas.

O nuevamente, en virtud de que yo tengo hermanos, soy capaz de tener hijos que sean sobrinos. Un gemelo molecular que no tenga hermanos no podría ipso facto tener hijos que sean sobrinos. Pero considero que es obvio a priori que la diferencia entre tener hermanos y no

tenerlos no constituye una diferencia en los poderes causales de los padres en virtud de su responsabilidad por esa diferencia entre las propiedades de sus descendientes.[9]

Mientras estemos en el tópico de qué es obvio a priori, los tres ejemplos precedentes podrían sugerir que la diferencia entre *PC*1 y *PC*2 en las instancias del esquema *S*, no constituyen una diferencia de los poderes causales cuando *PC*1 y *PC*2 son propiedades *relacionales*. Pero un momento de reflexión muestra que esto no puede ser correcto (como, ciertamente, se insistió en *Psychosemantics* hasta el cansancio). La taxonomía basada en propiedades relacionales es ubicua en la ciencia, y no está en discusión que propiedades como *ser un meteorito* o *ser un planeta* —propiedades que podrían distinguir, adviértase, pedazos de roca molecularmente idénticos— constituyan poderes causales. Es porque *esta* roca-gemela es un planeta y *esa* roca-gemela no lo es, que esta roca-gemela tiene una órbita kepleriana y esa roca-gemela no la tiene; es porque esta roca-gemela es un meteorito y esa roca-gemela no lo es, que los efectos de esta roca-gemela incluyen cráteres y los efectos de esa roca-gemela no los incluyen. Pero, evidentemente, *ser un planeta* y *ser un meteorito* son propiedades relacionales en un sentido pleno. Ser un planeta es ser una roca (o lo que fuere) que está girando alrededor de una estrella; ser un meteorito es ser una roca (o lo que fuere) que entra en colisión o que ha entrado en colisión con otra.

Puesto que las intuiciones son bastante fuertes en todos estos casos, hay razón para suponer prima facie que existen algunas condiciones de las propiedades-causa que son poderes causales, las cuales se pueden reconocer "desde un sillón": no todas las propiedades-causa son poderes causales; no todas las propiedades relacionales no lo son, en virtud de un par de ejemplos. Ahora enunciaré una condición que, afirmo, tiene que ser satisfecha si una propiedad de una causa es un poder causal en

9. Ésta es otra intuición taxonómica a priori que reclama ser tomada en serio. Es ridículo sugerir (como he señalado en *Psychosemantics*) que los estados neurológicos (o bioquímicos, o moleculares) deban ser taxonomizados haciendo referencia a las clases de propiedades que distinguen a los gemelos en los ejemplos estándares: si hay agua o gagua en las charcas locales, por ejemplo. Burge señala que quizá todo esto muestra que la taxonomía psicológica es más sensitiva, contextualmente, que la taxonomía neurológica. Pero, por cierto, esto no funciona: si alguien descubriera (por ejemplo) que vivir cerca de cables de alta tensión torna verdes a nuestras dendritas, podemos apostar a que nuestros neurólogos le prestarían atención. Así que, ¿cuál es la diferencia entre vivir cerca de cables de alta tensión y vivir cerca de charcas de gagua tal que uno, pero no el otro, es un candidato para ser un poder causal neurológico? (Ver más abajo).

virtud de ser responsable de una cierta propiedad de un efecto. Y alegaré que las propiedades que distinguen a los gemelos (como la de estar causalmente conectados a agua más que a gagua, o tener pensamientos [acerca] de agua más que pensamientos [acerca] de gagua) no satisfacen esta condición en virtud de su responsabilidad en las diferencias entre las propiedades intencionales de las conductas de los gemelos. Mi evidencia a favor de la aceptabilidad de esta condición será, en gran parte, que ubica ejemplos como los que acabo de presentar, de un modo intuitivamente satisfactorio.

Sin embargo, necesitamos limpiar el terreno antes de pararnos en él. (Le pido paciencia al lector; una vez que el terreno quede limpio seremos capaces de movernos muy rápido.) Para empezar quiero llamar la atención sobre lo que resulta ser una propiedad crítica del esquema *S*: la cuestión que estamos planteando no es si la diferencia entre tener *PC*1 y tener *PC*2 es una diferencia en los poderes causales; es, más bien, si la diferencia entre tener *PC*1 y tener *PC*2 es una diferencia en los poderes causales *en virtud de su ser responsable de una cierta diferencia entre E1 y E2*, a saber, en virtud de su ser responsable de que los *E*1 tengan *PE*1 en vez de *PE*2 y de que los *E*2 tengan *PE2* en vez de *PE*1. El punto que quiero enfatizar es que una propiedad-causa podría no contar como un poder causal en virtud de ser responsable de alguna propiedad-efecto, pero que podría constituir aun un poder causal en virtud de ser responsable de alguna otra propiedad-efecto.

El estar mi moneda de cara (en vez del estar de cruz) no constituye un poder causal en virtud de ser responsable de que el universo esté poblado con partículas H, pero podría constituir un poder causal en virtud de ser responsable del modo como mi moneda refleja la luz, en vez de hacerlo de algún otro modo (por ejemplo, del modo como lo haría si estuviese de cruz). Igualmente, mi tener hermanos no constituye un poder causal en virtud de que me permita tener hijos que sean sobrinos. Pero supongamos que hay una cosa tal como la enfermedad de tener un hermano. Ella causa que las personas que tienen hermanos padezcan una erupción cutánea. Tener un hermano, entonces, podría ser un poder causal en virtud de ser responsable de que las personas contraigan la enfermedad de tener un hermano. Insisto en este punto porque, si no se relativiza de este modo la cuestión acerca de los poderes causales, conduce a la suposición débil de que toda propiedad contingente es un poder causal. La suposición débil es que es (nomológicamente) posible construir un detector [*detector*] para cualquier propiedad contingente.

Considérese, por ejemplo, la propiedad de haber tenido una abuela

búlgara que arrancó un narciso cuando caminaba airosa hacia el mercado. Éste es el tipo de propiedad que puede distinguirte de tu gemelo molecular. Y, sin duda, tus efectos tienen propiedades por haber tenido tal abuela y los suyos no los tienen porque él no la tuvo; por ejemplo, todos y cada uno de tus efectos han adquirido la propiedad de ser los efectos de alguien que tuvo una abuela búlgara que..., etcétera. Pero intuitivamente, esto no es suficiente para hacer un poder causal del tener una A-búlgara..., o el no tener una A-búlgara que... En realidad, no habrías pensado que *cualquiera* de las propiedades de los efectos dependientes de tener una abuela búlgara —cualquiera de las propiedades que distinguen los efectos del gemelo molecular con una A-búlgara... de los efectos del gemelo molecular sin ella—, fueran del tipo correcto para hacer del tener una abuela búlgara un poder causal.

Y sin embargo sería posible aun construir (con más precisión, seguramente habría sido posible construir) una máquina que examine de modo exhaustivo la parte del espacio-tiempo que comienza con el nacimiento de tu abuela y termina con el tuyo, y que entre en un estado si detecta a alguien que fue tu abuela, que fue búlgara y arrancó un narciso cuando caminaba airosa hacia el mercado, pero que entra en un estado diferente en caso de que no detecte una persona tal. Supongo que es posible (en principio) construir una máquina que detecte confiablemente esa propiedad. Estoy dispuesto a suponer que es posible (en principio) construir una máquina tal para cualquier propiedad contingente, sea la que fuere. Si esta suposición es correcta, el detector de abuela búlgara puede distinguir entre tú y tu gemelo molecular, y tener una abuela búlgara es tener un poder causal *en virtud de los efectos (actuales o posibles) que las instanciaciones de esta propiedad tienen en los detectores de abuelas búlgaras*. (En general, si cualquier propiedad contingente puede ser detectada, entonces cualquier propiedad contingente es un poder causal en virtud de los efectos de sus instanciaciones sobre sus detectores.) Lo que no se sigue, sin embargo, es que el haber tenido una A-búlgara... sea un poder causal *en virtud de sus efectos sobre tu conducta*; o por cierto, en virtud de ser responsable de *cualquiera de tus propiedades*.

Respondo con estas palabras a un argumento sugerido por Burge (pero que él *no* apoya), que podría llevar a pensar que tener un pensamiento [acerca] de agua (en vez de un pensamiento [acerca] de gagua) tiene que ser un poder causal. Después de todo, ese argumento dice que tiene que ser posible, en principio, construir una máquina que examine la parte del espacio-tiempo que comienza con mi nacimiento y termina ahora, y que entre en un estado, si estoy conectado (de la manera

correcta) con el agua, pero que pase a otro estado si estoy conectado (de la manera correcta) con el gagua. Esta máquina responde de manera diferente a mi gemelo molecular y a mí, y lo hace en virtud del hecho de que estoy conectado con el agua del modo en que mi gemelo está conectado a XYZ. Así, mi duplicado y yo diferimos en nuestro poder de llevar a cabo [*to effect*] los estados de tal máquina. Así, tener pensamientos [acerca] de agua en vez de tener pensamientos [acerca] de gagua (a saber, tener el tipo de pensamiento que se tiene si se está conectado con agua más bien que con gagua) es un poder causal en virtud de su responsabilidad por estar la máquina en el estado en que está. Así, si el individualismo dice que tener pensamientos [acerca] de agua en vez de pensamientos [acerca] de gagua *no* es un poder causal, entonces es falso.

El individualismo *no* dice, sin embargo, que tener pensamientos [acerca] de agua en vez de pensamientos [acerca] de gagua no sea un poder causal. Lo que dice es que tener pensamientos [acerca] de agua en vez de pensamientos [acerca] de gagua no es un poder causal *en virtud de ser responsable de tu producción de conductas [referentes a] agua en vez de conductas [referentes a] gagua.* (E igualmente, estar conectado con agua en vez de gagua, no es un poder causal en virtud de ser responsable por tu tener pensamientos [acerca] de agua en vez de tener pensamientos [acerca] de gagua.) Esto equivale a decir que la diferencia entre tener pensamientos [acerca] de agua y tener pensamientos [acerca] de gagua, no es un poder causal en virtud de ser responsable de aquellas de tus propiedades *que son relevantes para las clases naturales psicológicas a las que tus pensamientos pertenecen*; por ejemplo, no es un poder causal en virtud de ser responsable de las propiedades de tu conducta.

Esto es todo lo que un individualista tiene que mostrar. Pedirle que muestre más sería hacer falso al individualismo si pudiera haber detectores para propiedades de contenido amplio. Pero esto sería trivializar la cuestión acerca del individualismo[10] ya que, como señalé antes, es plausible que pueda haber detectores para *cualquier* propiedad contin-

10. Aun cuando ser una partícula H fuera un poder causal en virtud de la posibilidad nomológica de construir detectores de partículas H. La detección de partículas es notoriamente cara y complicada. Pero una vez que se tiene un detector de partículas, resulta trivial convertirlo para detectar partículas H. Un detector de partículas H es un detector de partículas que está conectado a un adminículo que es sensible al anverso y reverso de mi moneda. Un detector que esté conectado a una abuela búlgara serviría para eso.

gente (y llegando al límite de las máquinas de Turing, también para un gran número de propiedades no contingentes).

Entonces: dos causas difieren en una cierta propiedad, y sus efectos difieren en una cierta propiedad, en virtud de esa diferencia en las causas; y nosotros deseamos saber cuándo la diferencia entre los efectos hace de la diferencia entre las causas un poder causal. En un momento lo diré. Pero puede tener valor enfatizar que los casos en los que estamos interesados no son aquellos en los que la propiedad que distingue las causas sea ella misma la propiedad de tener un cierto poder causal.

Considérese el caso en el que *PC*1 es la propiedad de tener el poder causal para producir eventos que tengan *PE*1. Entonces, por supuesto, la diferencia entre tener *PC*1 y no tenerla es la diferencia entre tener un cierto poder causal y no tenerlo. Este tipo de caso es fácil porque *no es contingente* que tener *PC*1 sea tener un poder causal. Hay otros casos de este tipo, menos transparentes. Por ejemplo, no es contingente que *ser soluble en agua* sea un poder causal, porque no es contingente que las cosas que tienen esa propiedad tengan el poder de disolverse en agua; y la propiedad de *ser un árbol de levas* es un poder causal porque no es contingente que las cosas que tienen esa propiedad tengan el poder para subir las válvulas en un cierto tipo de motor (dadas condiciones óptimas de funcionamiento..., etcétera). Hay, por supuesto, cuestiones científicamente interesantes acerca de las propiedades que son poderes causales no contingentes; por ejemplo, hay interesantes cuestiones acerca de la base de superveniencia de estas propiedades y acerca de su análisis adecuado en términos de microfunciones. Pero no hay cuestiones interesantes acerca de si las cosas que tienen esas propiedades tienen poderes causales en virtud de tenerlas. Esa cuestión se responde a sí misma.

Compárense, no obstante, las propiedades de ser un planeta, ser un meteorito, y similares. Ellas son poderes causales en virtud de, por ejemplo, sus respectivas capacidades [*abilities*] para producir órbitas keplerianas y cráteres. Pero, en la medida en que ser un planeta sea un poder causal en virtud de la capacidad de los planetas de producir órbitas keplerianas, es contingente que ser un planeta sea un poder causal (porque es contingente que los planetas tengan órbitas kleperianas); y en la medida en que ser un meteorito sea un poder causal en virtud de la habilidad de los meteoritos de hacer cráteres, resulta contingente que ser un meteorito sea un poder causal (porque es contingente que los meteoritos hagan cráteres).

Adviértase que los casos de contenido amplio son similares a los

casos del meteorito y del planeta, y diferentes de los casos disposicional y funcional. Pudiera ser que estar relacionado con agua en vez de gagua (y por lo tanto, tener pensamientos [acerca] de agua en vez de pensamientos [acerca] de gagua) sea tener un poder causal; pero si así fuese, es contingente que lo sea. La propiedad de estar conectado al agua no es *idéntica* a la propiedad de tener un cierto poder causal, aunque podría ser que hubiera poderes causales que uno tendría si uno estuviera relacionado con agua, y que uno no tendría si no lo estuviera.

Extraigo dos moralejas. Primero (en oposición a algunas sugerencias de R. Van Gulick),[11] el hecho de que tener una propiedad funcional sea ipso facto tener un cierto poder causal, no arroja una luz particular, de un modo u otro, sobre la cuestión de si tener una propiedad de contenido amplio es tener un poder causal. Esta conclusión puede parecer paradójica ya que, después de todo, se supone que esas propiedades psicológicas *son* funcionales; así, si las propiedades funcionales son poderes causales de manera no contingente, y si las propiedades psicológicas son funcionales, ¿cómo podrían los pensamientos [acerca] de agua y los pensamientos [acerca] de gagua dejar de ser poderes causales?

La respuesta es que realmente no está en disputa si los pensamientos [acerca] de agua y los pensamientos [acerca] de gagua son poderes causales. Todo lo contrario, *por supuesto* que lo son: mis pensamientos [acerca] de agua son causalmente responsables de que yo busque agua; los pensamientos de mi gemelo [acerca] de gagua son causalmente responsables de su búsqueda de gagua..., y así sucesivamente. La cuestión sobre la cual gira la superveniencia local —y por lo tanto, el individualismo— es, no obstante, si la diferencia entre los pensamientos [acerca] de agua y los pensamientos [acerca] de gagua es una *diferencia* en poderes causales. El antiindividualista dice 'Sí, lo es, en virtud de la diferencia intencional entre las conductas que causan los pensamientos [acerca] de agua y los pensamientos [acerca] de gagua'. El individualista dice 'No, no lo es. Los pensamientos [acerca] de agua y los pensamientos [acerca] de gagua tienen los mismos poderes causales, sólo que están instanciados en personas con diferentes historias causales'. La idea de que los estados mentales son roles funcionales resuelve esta discusión sólo en base a la presuposición, afectada por una petición de principio, de que las conductas [referentes a] agua y las conductas [referentes a] gagua, son conductas de diferentes clases.

11. "Metaphysical Arguments for Internalism and Why They Don't Work", en S. Silvers (comp.), *Representations* (Philosophical Studies Series, 40). Boston, Kluwer, 1989.

Imagínese que tuviéramos diferentes palabras para *tener sed y nacer en Bronx* y *tener sed y nacer en Queens*. Entonces no estaría en discusión que tener sed en Bronx y tener sed en Queens, son poderes causales, tener sed en Bronx y tener sed en Queens, hacen ambas, por ejemplo, que la gente beba. Pero podría plantearse la cuestión de si tener sed en Bronx y tener sed en Queens son poderes causales *diferentes*.[12] Mi tesis es que esa cuestión quedaría abierta, *aun cuando* no esté en discusión que tener sed en Bronx y tener sed en Queens sean poderes causales.

Un funcionalista podría proponerse argumentar que son poderes causales diferentes porque, de un lado, los estados psicológicos son individuados funcionalmente, y porque, del otro lado, la sed en Bronx y la sed en Queens conducen a diferentes consecuencias conductuales, es decir, a conductas de aplacar la sed en Bronx, en un caso, y a conductas de aplacar la sed en Queens, en el otro. Pero es claro que argumentar de este modo sería caer en una petición de principio. Porque quienquiera que niegue que la sed en Queens y la sed en Bronx son poderes causales diferentes, negará también —y por la misma razón— que la conducta de aplacar la sed en Bronx y la conducta de aplacar la sed en Queens, son conductas de diferentes clases. Por lo tanto, el principio de que los estados psicológicos son individuados funcionalmente, no resuelve en estos casos las cuestiones acerca de la identidad y la diferencia de los poderes causales. Y, por supuesto, para un individualista, 'Quiere agua' *es* esa clase de caso; para un individualista, 'Quiere agua' significa algo así como "Sediento y nacido *aquí*".

Lo que distingue a los gemelos y amenaza la superveniencia local es la propiedad de *tener pensamientos* [*acerca*] *de agua en vez de tener pensamientos* [*acerca*] *de gagua*. Sin duda, si el funcionalismo es verdadero, entonces los estados mentales son todos ellos poderes causales. Pero, por supuesto, no se sigue del funcionalismo que las diferencias entre los estados mentales sean todas ellas diferencias funcionales; no se sigue entonces que las diferencias entre los estados mentales sean todas ellas diferencias de poder causal. De hecho, hasta donde puedo ver, la única conexión relevante entre el funcionalismo y los temas presentes acerca del individualismo es ésta: si como he alegado, la diferencia entre los estados mentales de los gemelos sólo en el mejor de los casos es *contingentemente* una diferencia de poderes causales, se sigue

12. Esto bien podría resultar ser la misma cuestión que si 'está sediento-en-Bronx' y 'está sediento-en-Queens' son proyectables; si, por ejemplo, las generalizaciones psicológicas acerca de la gente sedienta en Bronx, como tal, son confirmadas por sus instancias.

que *no puede* ser una diferencia de rol funcional, dado que las diferencias de rol funcional son diferencias de poderes causales *no* contingentemente.[13] Compárese ser soluble en agua en vez de soluble en gagua. *No* es contingente que tener *esta* propiedad sea tener un poder causal; ser soluble en agua pero no en gagua es justamente tener el poder de disolverse en la primera pero no en la segunda. Es correcto conceder, entonces, que *ser soluble en agua en vez de ser soluble en gagua* es una propiedad funcional, como, por cierto, demanda la intuición.

Mi impresión es que, razonablemente, todo esto no es tendencioso. En general, los amigos del contenido amplio argumentan que muy bien *podría resultar* —que no hay ninguna razón metafísica por la que no debiera resultar—, que tener estados mentales que difieren en su contenido amplio es tener estados mentales que difieren en sus poderes causales. (Por ejemplo, pueda ser que haya leyes causales que distingan entre gemelos.) Pero no recuerdo haber oído que alguien argumente que tener estados mentales que difieren en el modo en que lo hacen los estados mentales de los gemelos, *sea* tener estados mentales que difieren en sus poderes causales. Por el contrario, de acuerdo con lo que habitualmente se entiende, son las *historias* causales de los gemelos las que distinguen sus estados mentales. Y la intuición que hay acerca de los rasgos de la historia causal es que algunos de ellos son poderes causales (por ejemplo, *haberse atascado en el tránsito*; *haber sido vacunado contra la varicela*) y algunos no lo son (por ejemplo, *haber tenido una abuela búlgara*; *haber nacido en jueves*), y es contingente cuáles son cuáles.

La segunda moraleja que extraigo, entonces, es que en nuestra búsqueda de una condición útil respecto de qué cosa hace de una diferencia de las propiedades causales una diferencia de los poderes causales, podemos restringirnos a casos donde es *contingente* que la diferencia en las propiedades constituya una diferencia entre poderes. Estamos llegando finalmente al punto álgido.

Ahora voy a contar, por fin, la historia [*story*] de por qué ser un planeta (por ejemplo) es un poder causal y tener hermanos (por ejemplo) no lo es. Propongo hacer esto en dos pasos. Primero daré una versión

13. Esto concuerda con la idea de que, mientras que *tener una creencia* (un deseo, o lo que se quiera) es estar en el estado funcional correcto, tener una creencia *que P* es cuestión de tener relaciones correctas cabeza-mundo. Así, la diferencia entre tener una creencia que P y una creencia que Q *no es* una diferencia funcional. Más sobre esto en *Psychosemantics*.

simplificada de la historia: tiene la virtud de volver relativamente transparente la idea básica, pero tiene la desventaja de que no funciona. Haré entonces, los movimientos técnicos que se requieran para taponar el agujero; la versión complicada que emerge será una condición motivada sobre la propiedad de una causa que es un poder causal; una condición que, afirmo, las propiedades del contenido amplio no cumplen.[14]

Y bien, aquí estoy yo y aquí está mi gemelo molecular; yo tengo hermanos y él no; en virtud de mi tener hermanos mis hijos son sobrinos, y en virtud de su no tener hermanos, sus hijos no lo son; y lo que queremos saber es: ¿por qué *tener hermanos* no es un poder causal en virtud de ser responsable de esta diferencia en nuestra descendencia? He aquí un primer ensayo de respuesta: es porque tener hermanos está *conceptualmente* conectado con el tener hijos que son sobrinos; ser un sobrino *es justamente* ser un hijo cuyos padres tienen hermanos. Y, en términos aproximados, sus poderes causales son una función de sus conexiones *contingentes*, no de sus conexiones conceptuales. Tal como, por cierto, nos enseñó el tío Hume.

De modo similar: aunque es un hecho que todas las partículas del mundo se vuelven partículas H cuando mi moneda está de cara, ese hecho no hace del estar de cara un poder causal de mi moneda. Eso es porque la conexión entre todas las partículas del mundo que se vuelven partículas H en *t* y el estar de cara de mi moneda en *t*, es conceptual. Ser una partícula H en *t es justamente* ser una partícula en el instante en el que mi moneda está de cara.

Compárense los casos de propiedades relacionales que son realmente poderes causales, como *ser un planeta*. Ser un planeta es un poder causal en virtud de, por ejemplo, su conexión contingente (*a fortiori*, no conceptual) con tener una órbita kepleriana. Esto es, ser un planeta es un poder causal porque es verdadero y contingente que si se tienen trozos de rocas molecularmente idénticos, uno de los cuales es un planeta y el otro no lo es, entonces, *ceteris paribus*, lo que es un planeta tendrá una órbita kepleriana, y *ceteris paribus*, lo que no es un planeta, no la tendrá.

14. Quizá sea mejor reiterar que esto es una abreviatura de: "...una condición motivada para que una diferencia entre propiedades de las causas sea una diferencia en sus poderes causales", y que la afirmación no es que ser un pensamiento [acerca] de agua (/de gagua) no sea un poder causal, sino que la diferencia entre tener un pensamiento de agua y tener un pensamiento de gagua no es una diferencia en poder causal. Más generalmente, la afirmación es que ninguno de los dos estados difiere en sus poderes causales *justamente* en virtud de diferir en sus contenidos amplios.

He aquí la forma general de la solución propuesta.[15] Considérese una instancia del esquema *S*. *C*1 tiene *PC*1, *C*2 tiene *PC*2, *E*1 tiene *PE*1, *E*2 tiene *PE*2, y la diferencia entre las causas es responsable de la diferencia entre los efectos, en el sentido de que *E*1 no habría tenido *PE*1 (en vez de *PE*2) y que *C*1 tuvo *PC*1 (en vez de *PC*2). Y lo que queremos saber es: ¿cuándo el hecho de que esa diferencia en las causas es responsable de esa diferencia en los efectos, hace de *PC*1 y *PC*2 poderes causales? La respuesta, que llamaré *condición C*, es:

> sólo cuando no es una verdad conceptual que las causas que difieren en que una tiene *PC*1 cuando la otra tiene *PC*2, tienen efectos que difieren en que una tiene *PE*1 cuando la otra tiene *PE*2.[16]

Así, por ejemplo, que *ser un meteorito* sea un poder causal en virtud del hecho de que los meteoritos son responsables de los cráteres, concuerda con la condición *C*. Tómese un par de rocas gemelas tales que una es un meteorito y la otra no. Entonces (*ceteris paribus*) los cráteres estarán entre los efectos de la primera pero no entre los efectos de la segunda. Y la relación entre la diferencia que se da entre las rocas gemelas y la diferencia que se da entre sus efectos, es no conceptual. Por lo tanto, está bien que ser un meteorito (en vez de ser un meteorito gemelo) resulte ser un poder causal en virtud del hecho de que los meteoritos causan cráteres y los meteoritos gemelos no los causan.[17]

15. *Por favor*, téngase en cuenta que esta solución vale para los casos en los que, si las propiedades en consideración son poderes causales, entonces son poderes causales de manera contingente. En particular, *no* se supone que valga para las propiedades que son poderes causales de manera necesaria, como las disposiciones. Ver arriba.

16. Quiero enfatizar que la condición *C* no impugna la doctrina davidsoniana de que la necesidad/contingencia de las relaciones entre *eventos* es relativa a la descripción [*description-relative*]. Supóngase que *e*1 cause *e*2, de modo que '*e*1 causa *e*2' es verdadera de manera contingente. Sin embargo, habrá descripciones satisfechas por *e*1 que implican (/presuponen) que *e*1 causa *e*2 (por ejemplo, descripciones tales como 'la causa de *e*2'. 'La causa de *e*2 causó *e*2' es, por supuesto, necesariamente verdadera). Pero nada de esto implica que haya algo relativo a la descripción acerca de si la instanciación de una propiedad implica la instanciación de otra (acerca de si, por ejemplo, es conceptualmente necesario que lo que instancie la propiedad de ser soltero instancia la propiedad de ser no casado; o si es conceptualmente necesario que todo lo que instancia *pensar* [*acerca*] *de agua* instancie *estar causalmente conectado con agua*. Es esta última clase de afirmaciones, no la primera, la que está en cuestión respecto de la condición *C*.

17. Por contraste, considero que la capacidad para producir cráteres [de] *meteorito* (donde cráter [de] meteorito es un cráter causado por un meteorito) no es un poder causal que tienen los meteoritos por encima y por debajo de su poder de producir cráteres *tout*

Nótese que la moraleja es que está bien que ser un meteorito sea un poder causal en virtud de ese hecho; no que ser un meteorito *es* un poder causal en virtud de ese hecho. Satisfacer la condición *C* es, supongo, necesario pero no suficiente para ser un poder causal. No obstante, como veremos, las propiedades de contenido amplio *no* satisfacen esa condición necesaria, y ello es suficiente para reivindicar al individualismo.

Considérese, primero, la propiedad de tener en su historia agua (la propiedad de estar conectado [*connected*] a agua de un modo como yo lo estoy y mi gemelo no). La diferencia entre estar conectado así y no estar conectado así, es responsable de una cierta diferencia entre los contenidos amplios de mis pensamientos y los contenidos amplios de mi gemelo, a saber, que yo tengo pensamientos [acerca] de agua y él tiene pensamientos [acerca] de gagua. Lo que queremos saber es: ¿cuenta como un poder causal esta diferencia entre nuestras historias, en virtud de la diferencia entre los contenidos de nuestros pensamientos de la que es responsable? Y la respuesta es: "No, porque es *conceptualmente necesario* que si se está conectado a agua del modo correcto, entonces se tienen pensamientos [acerca] de agua (en vez de pensamientos [acerca] de gagua), y es conceptualmente necesario que si se está conectado a gagua del modo correcto entonces se tienen pensamientos [acerca] de gagua (en vez de pensamientos [acerca] de agua)". Tener un pensamiento [acerca] de agua es justamente tener un pensamiento que está conectado con agua del modo correcto, y tener un pensamiento [acerca] de gagua es justamente tener un pensamiento que está conectado con gagua del modo correcto.[18] Por lo tanto no es el caso que mi estar conectado a agua en vez de estar conectado a gagua, sea una diferencia en mis poderes causales en virtud de ser responsables de mi tener pensamientos [acerca] de agua en vez de pensamientos [acerca] de gagua.

---

court. Esto se debe al tipo de razón conocida: es conceptualmente necesario que, mientras que un cráter causado (del modo correcto) por un meteorito es un cráter [de] meteorito, un cráter causado (de cualquier manera) por un no meteorito molecularmente idéntico, no lo es. Análogamente la capacidad de producir quemaduras solares no es un poder causal que el sol tenga por encima y por debajo de su poder de producir quemaduras.

18. Más precisamente (y suponiendo que el funcionalismo tiene razón respecto de qué son las creencias, los deseos, y cosas similares), tener un pensamiento [acerca] de agua es tener un pensamiento que (a) está conectado con el agua del modo correcto; y (b) tiene las propiedades funcionales que comparten los pensamientos [acerca] de agua y los pensamientos [acerca] de gagua. Dejo a (b) fuera del texto para simplificar la exposición. Sin duda, la exposición podría ocuparse de algo de eso.

Considérese ahora la diferencia entre tener pensamientos [acerca] de agua y tener pensamientos [acerca] de gagua. Los pensamientos [acerca] de agua causan conducta [referida a] agua (perforar para obtener agua, y demás); los pensamientos [acerca] de gagua causan conducta [referida a] gagua (perforar para obtener gagua, y demás). Lo que queremos saber es: ¿cuenta la diferencia entre tener pensamientos [acerca] de agua y tener pensamientos [acerca] de gagua, como un poder causal en virtud del hecho de ser responsable de esa diferencia en las propiedades intencionales de la conducta del que piensa? Y la respuesta es: "No, porque es conceptualmente necesario que la gente que posee pensamientos [acerca] de agua (en vez de pensamientos [acerca] de gagua) produzca conducta [referida] a agua (en vez de conducta [referida a] gagua)". Ser una conducta [referida a] agua (en vez de una conducta [referida a] gagua) es ser una conducta que es causada por pensamientos [acerca] de agua (en vez de pensamientos [acerca] de gagua). Así, aunque sea verdad que los pensamientos [acerca] de agua son responsables de la conducta [referida a] agua, mientras que los pensamientos [acerca] de gagua no lo son, no se sigue que los pensamientos [acerca] de agua tengan un poder causal que los pensamientos [acerca] de gagua no poseen. Por el contrario, ser alguien que piensa en agua es el mismo poder causal que ser alguien que piensa en gagua, sólo que instanciado en una persona con una historia causal diferente.

He estado diciendo que sólo cuando la diferencia entre las causas no está conceptualmente conectada con las diferencias correspondientes en los efectos, la diferencia en las causas cuenta como una diferencia en los poderes causales. Pienso que esto es realmente el núcleo de la cuestión; por qué ser responsable de las partículas H, ser responsable de hijos que son sobrinos, ser responsable de la conducta que es una conducta [referida a] agua, y otros casos similares, no cuentan como poderes causales. Pero cuidado, la propuesta no funciona como está establecida y tendré que hacer algo para remendarla.

Supóngase que el agua es la bebida favorita de Bush. Entonces tenemos:

1. Si estoy conectado con agua de la manera correcta, entonces mis pensamientos son pensamientos [acerca] de agua,

y

2. Si estoy conectado con agua de la manera correcta, entonces mis

pensamientos son pensamientos acerca de la bebida favorita de Bush.

Adviértase que tanto 1 como 2 me distinguen de mi gemelo: dado que él no está conectado con agua de la manera correcta, es falso que tenga pensamientos [acerca] de agua, y también es falso que sus pensamientos sean acerca de la bebida favorita de Bush. Y, aunque 1 es conceptualmente necesaria, 2 es contingente. Es decir, aun cuando la diferencia entre estar conectado a agua y estar conectado a gagua, no satisface la condición *C* por ser responsable de la diferencia entre tener pensamientos [acerca] de agua y tener pensamientos [acerca] de gagua, *satisface* la condición *C* por ser responsable de tener pensamientos que son acerca de la bebida favorita de Bush y tener pensamientos que no lo son. De tal modo, satisface la condición *C* en virtud de producir alguna diferencia prima facie *psicológica*, alguna diferencia que es prima facie relevante para la taxonomía psicológica. Así, aun cuando, como afirmo, la condición *C* tiene su núcleo en el lugar correcto, sin embargo, tal como se la enunció, no tiene "garra". ¡Lástima! [19]

Sin embargo, todo va a andar bien. Sea *G* una propiedad que tienen los sobrinos, que sea tan contingente como usted quiera. (*G* podría ser la propiedad de brillar en la oscuridad, si es que los sobrinos hacen eso.) Entonces, es necesario:

3. Si *G* es una propiedad que tienen los sobrinos, entonces si tengo hermanos, mis hijos tienen *G*.

Así, por ejemplo, es necesario que (si los sobrinos brillan en la oscuridad, entonces, si tengo hermanos, mis hijos brillan en la oscuridad). La razón por la cual esto es necesario es, en términos generales, que es conceptualmente necesario que si usted tiene un hermano, sus hijos tienen cualquiera de las propiedades que tienen los sobrinos. Adviértase, para repetirlo, que esto es verdadero aun respecto de las propiedades que los sobrinos tienen de manera contingente: es conceptualmente necesario que si *P* es una propiedad que los sobrinos tienen de manera contingente, entonces si usted tiene hermanos, sus hijos tienen *P*. (La que *no* es verdadera, por supuesto, es 3', la variante de 3 que tiene un operador

19. Estoy en deuda con Stich por esta línea argumentativa. Supongo que piensa que debería estar agradecido.

modal en su interior. 3' es falsa en los casos en los que *P* es una propiedad que los sobrinos tienen de manera contingente.)

3'. Si *P* es una propiedad de los sobrinos, si usted tiene hermanos, es necesario que sus hijos tengan *P*.

Ahora considérese 4, que para mí es claramente contingente.

4'. Si *E* es una propiedad que tienen las órbitas keplerianas, si soy un planeta, mi órbita tiene *E*.

(Uno podría considerar 4 leyendo a *E* como la propiedad de ser elíptica.)

La razón de que 4 sea contingente es, aproximadamente, que es contingente que los planetas tengan órbitas keplerianas, de modo que es contingente que si voy a un planeta, entonces mi órbita tiene las propiedades keplerianas, que las órbitas tienen. Esto es verdad aun si *E* es una propiedad que las órbitas keplerianas tienen necesariamente (como la de ser una órbita kepleriana).[20]

Ahora podemos ver el caso que antes nos molestaba. Considérese 5, donde *B* podría ser la propiedad de tener un pensamiento acerca de la bebida favorita de Bush.

5. Si *B* es una propiedad que tienen los pensamientos [acerca] de agua, entonces, si estoy conectado con el agua de manera correcta, *B* es una propiedad que tienen mis pensamientos.

Considero que 5 es, de manera obvia, conceptualmente necesaria. La razón, aproximadamente, es que es conceptualmente necesario que si estoy conectado con agua de la manera correcta, entonces, mis pensamientos son pensamientos [acerca] de agua, y es una verdad trivial que si algo es una propiedad de los pensamientos [acerca] de agua, es una propiedad de mis pensamientos, si mis pensamientos *son* pensamientos [acerca] de agua.

De manera semejante, mutatis mutandis, con 6.

20. Sin embargo, 4 es necesaria si usted elige *E* como una propiedad que *todo* tiene de manera necesaria, como ser auto-idéntico, por supuesto. Esto no prejuzga acerca del presente punto.

6. Si *B* es una propiedad que tienen las conductas [referidas a] agua, entonces, si mis pensamientos son pensamientos [acerca] de agua, mis conductas tienen *B*.

6 es conceptualmente necesaria porque, aproximadamente, es conceptualmente necesario que las conductas que causan los pensamientos [acerca] de agua, sean conductas [referidas a] agua; por lo tanto, es conceptualmente necesario que si algo es una propiedad de las conductas [referidas a] agua, entonces es una propiedad de las conductas de los que piensan en agua. Esto es verdad aun si *B* es una propiedad que las conductas [referidas a] agua tienen de manera contingente, tal como ser conductas que tienen que ver con la bebida favorita de Bush. [21]

Compárese estos casos con otros en los que los pensamientos tengan poderes causales genuinos. Supóngase que pensar en la topología le produce a uno dolor de cabeza. Entonces tenemos 7:

7. Si *B* es una propiedad de los dolores de cabeza, si tengo pensamientos [acerca de la] topología, mi estado mental tiene *B*.

7 es claramente contingente, y lo es debido a la contingencia de la relación (putativa) entre algo que es un pensamiento [acerca de la] topología y tener entre sus efectos dolores de cabeza. Adviértase que 7 es contingente aun si *B* es una propiedad que poseen necesariamente los dolores de cabeza (tal como ser dolores de cabeza).

Por lo tanto, aquí está la historia. Para que la diferencia entre ser *PC*1 y ser *PC*2 sea una diferencia de poderes causales, tiene que darse, al menos, que los efectos de ser *PC*1 difieran de los efectos de ser *PC*2. Pero, afirmo, se requiere además que esta diferencia entre los efectos esté relacionada de manera *no conceptual* con la diferencia entre las causas. Esta condición adicional está motivada por nuestras intuiciones acerca de los ejemplos y por la tesis humeana de que los poderes causales son, después de todo, poderes que entran en relaciones no conceptuales. Sin embargo, las diferencias en el contenido amplio no satisfacen

21. Para quienes gustan de las minucias técnicas: en efecto, la línea de mi argumentación original dependía de que hubiera una relación conceptual entre la conexidad con agua [*water connectedness*] y los pensamientos [acerca] de agua interpretados *de dicto*. Stich obtiene una relación *contingente* entre la conexidad con agua y los pensamientos *de re* [acerca] de agua al invocar premisas contingentes tales como, por ejemplo, que el agua es la bebida favorita de Bush; así satisface trivialmente la condición *C*. En efecto, 5 proporciona una versión de *C* al condicionalizar esas premisas contingentes.

per se esta condición. Hay diferencias entre mi conducta y la de mi gemelo que se deben, en primera instancia, a la diferencia entre los contenidos intencionales de nuestros pensamientos y, en segunda instancia, a mi estar conectado con agua de un modo en que mi gemelo no lo está. Pero esas diferencias entre los efectos están relacionadas conceptualmente con las diferencias entre las causas; es conceptualmente necesario que el estar conectado con agua en vez de estar conectado con gagua, lleve a pensar en agua y no en gagua; y es, otra vez, conceptualmente necesario que pensar en agua conduzca a conductas [referidas a] agua y el pensar en gagua no lo haga.

Así, pues, la diferencia entre los estados mentales de los gemelos no cuenta como una diferencia en el poder causal en virtud de su responsabilidad por las diferencias intencionales entre las conductas de los gemelos. Por lo tanto, el argumento *B* no es bueno; lo que está mal en él es que la inferencia de 2' a 3' es errónea.[22] Finalmente, dado que se supone que los efectos de los estados mentales que difieren sólo en el contenido amplio son (de manera relevante) diferentes *sólo* bajo una

22. Puedo imaginar que alguien podría ahora decir: "Todo esto se sigue por haber concedido que las propiedades intencionales que distinguen a los gemelos son poderes causales de manera contingente, si son poderes causales. Bien, estaba en un error: aquellos tipos de diferencia entre los estados con contenido amplio, son diferencias *no* contingentes de poderes causales. Por ejemplo, la propiedad de tener pensamientos [acerca] de agua (y no [acerca] de gagua) es *idéntica* a la propiedad de poder producir conductas [referidas a] agua (y no [referidas a] gagua); la propiedad de estar conectado con agua (y no con gagua) es idéntica a la propiedad de poder tener pensamientos [acerca] de agua (no [acerca] de gagua), etcétera. Adviértase que ser soluble es un poder causal y no está conectado conceptualmente con la diferencia entre disolverse y no disolverse".

Es correcto, pero realmente no ayuda. Si tener pensamientos [acerca] de agua (y no [acerca] de gagua) es idéntico a tener (entre otras) al poder de buscar agua (y no gagua), entonces las generalizaciones psicológicas de contenido amplio que distinguen a los gemelos [como 'Si usted tiene deseos [acerca] de agua (y no [acerca] de gagua), entonces busca agua (y no gagua), resultan ser todas en sí mismas conceptualmente necesarias (un tipo de observación respecto de la cual los ryleanos estaban, por supuesto, muy alertas). Por lo tanto, si la suposición es que los estados intencionales son poderes causales de manera no contingente, entonces la forma apropiada de la alegación de superveniencia que hace el individualista es que ninguna generalización intencional *contingente* puede distinguir [entre] gemelos (ninguna generalización intencional contingente puede ser tal que uno pero no el otro de los gemelos satisface su antecedente o su consecuente). La moraleja sería entonces: no hay leyes causales acerca de los estados intencionales como tales (de hecho, ésa es justo la clase de moraleja que los ryleanos acostumbraban hacer).

Es verdad que ser soluble es un poder causal aun si está conceptualmente conectado con disolverse. Pero el precio por evadir así la condición *C* es el *status* "quasi-lógico" de 'Si es soluble, entonces se disuelve'.

descripción intencional, se sigue que *no* hay diferencias taxonómicamente relevantes que sean consecuencias respecto de las diferencias de contenido amplio en tanto tales. Desde el punto de vista de la taxonomía psicológica, mis estados mentales tienen que pertenecer, en consecuencia, a la misma clase natural que los de mi gemelo molecular. Así, el individualismo es verdadero y la superveniencia local se preserva. Final de la historia.

— ¡Oh! ¡Alto, ladrón! Detengan a esa persona.
— ¿Qué sucede?
— Me prometió un argumento a favor del contenido estrecho pero todo lo que me ha dado es un argumento *contra* argumentos en contra del individualismo. Quiero que me devuelva el dinero.

Hemos visto que los pensamientos [acerca] de gagua y los pensamientos [acerca] de agua no son poderes causales diferentes. Por lo tanto, a los fines del psicólogo, ellos son el mismo estado intencional.[23] Pero no pueden ser el mismo estado intencional a menos que tengan el mismo contenido intencional. Y no pueden tener el mismo contenido intencional, a menos que el contenido intencional sea individuado de manera estrecha. Éste es, ahora, un argumento a favor del contenido estrecho.

— Muchas gracias.
— De nada.

TRADUCTORES: Eduardo Barrio, Patricia Brunsteins, Margarita Roulet y Julia Vergara.
REVISIÓN TÉCNICA: Eduardo Rabossi.

23. —Pero, ¿por qué puede este psicólogo putativo no permitirles ser estados *diferentes* pero con los *mismos* poderes causales?
—Porque, si lo hace, su historia pierde las generalizaciones, a saber, todas las generalizaciones que me subsumen a mí y a mi gemelo. Una buena taxonomía lo es, acerca de generalizaciones que *no* se pierden.

VI

LOS MODELOS COMPUTACIONALES: LAS DISPUTAS DEL CONEXIONISMO Y LA INTELIGENCIA ARTIFICIAL

Capítulo 13

UNA INTRODUCCIÓN AL CONEXIONISMO *

John L. Tienson

La metáfora del computador ha jugado un rol prominente en la filosofía de la mente, al menos desde el advenimiento de la inteligencia artificial en la década del 50. El conexionismo es un nuevo desarrollo, algunos dicen una moda, en la ciencia cognitiva de la inteligencia artificial. Naturalmente, los filósofos están interesados en la pregunta de si el conexionismo puede proveer un modelo fructífero de la mente-cerebro, y en la pregunta relacionada de cuáles serán las implicaciones concernientes a la cognición, si el conexionismo es capaz de proveer tal modelo...

Puesto que el conexionismo aún no es familiar para muchos filósofos, parece apropiado empezar con una introducción breve, no técnica, a él. La reciente popularidad del conexionismo es en parte una respuesta a una crisis kuhniana en la inteligencia artificial. Comenzaré con una breve descripción de esa crisis.

1. La buena y anticuada inteligencia artificial

La inteligencia artificial (IA), como especialidad, se propone lograr que los computadores se comporten de manera que podamos reconocerlos como inteligentes. Una porción significativa de esa investigación tiene por objeto entender la inteligencia *natural* y, así, pretende lograr que los computadores se comporten de manera inteligente, en algún sentido *en que nosotros lo hacemos*.[1] La característica definitoria de los

* "An Introduction to Connectionism", *The Southern Journal of Philosophy* 126, (supl.), (1987), págs. 1-16. Con autorización del autor y del *Southern Journal of Philosophy*.

1. Por supuesto, debemos comenzar por donde podemos, con modelos simplificados de partes artificialmente aisladas de lo que hacemos, con la esperanza de que esto nos llevará a modelos menos simplificados de partes menos aisladas. Pero el objetivo es entender

computadores corrientes es que manipulan símbolos en términos de reglas precisas, que son manipuladores de símbolos gobernados por reglas. Contienen estructuras de datos —objetos estructurados sintácticamente, oraciones— y reglas que se refieren a esas oraciones. Un programa es una secuencia de tales reglas. Puesto que las reglas, en sí mismas, son también estructuras de datos, los programas pueden ser almacenados en la memoria para servir como rutinas reiterables accesibles por otros programas. La gracia de hacer un computador consiste en lograr que los procesos causales del mecanismo reflejen los procesos sintácticos especificados en los programas. Para que esto sea posible, las reglas tienen que ser precisas y sin excepciones.

La característica definitoria de la inteligencia artificial clásica, lo que John Haugeland denominó BAIA (la buena y anticuada inteligencia artificial),[2] es que la arquitectura del computador clásico llega a la esencia de la inteligencia: la cognición es lo que el computador clásico hace, manipulación de símbolos gobernados por reglas. El objetivo es escribir programas que resulten en un comportamiento inteligente, y finalmente escribir *programas* como aquellos que subyacen a nuestro comportamiento inteligente. Puesto que nosotros, los humanos, nos comportamos inteligentemente, se asume que tales programas pueden ser descubiertos.

En nuestro uso o comprensión de los programas, interpretamos las oraciones de los lenguajes computacionales. De tal modo, tienen contenido semántico, representan, aunque más no sea de segunda mano. BAIA asume que nosotros, los humanos, tenemos que tener representaciones comparables.[3] Así, el punto de vista clásico de la mente se denomina a menudo, el punto de vista de "reglas y representaciones", caracterizado por

1) Representaciones estructuradas sintácticamente.

---

cómo lo hacemos y en consecuencia un (vago) *desideratum* de los modelos es que resuelvan problemas de maneras que son similares de modo relevante, a las maneras en que nosotros resolvemos esos problemas.

2. John Haugeland, *Artificial Intelligence: The Very Idea* (Cambridge: Bradford Books, MIT Press, 1985).

3. Las cuales presumiblemente representan, no por interpretación, sino por naturaleza. Adviértase que en sus supuestos, BAIA no pretende *simular* la inteligencia, como simulamos huracanes y economías. Estar satisfecho con la simulación es abandonar el punto de vista de BAIA de que la cognición *está* ejecutando la clase de programa correcto.

2) Reglas formales, precisas que refieren a la estructura sintáctica de esas representaciones.[4]

2. *La crisis kuhniana en BAIA*

Existe una crisis kuhniana en BAIA ocasionada por un patrón de promesas incumplidas y de resultados decepcionantes. A lo largo de la historia de la inteligencia artificial, los investigadores han predicho con optimismo: "En cinco (o diez) años, los computadores harán esto y aquello: traducirán, o comprenderán el habla, o serán tan buenos expertos en...". Esto nunca ocurrió y, en general, ni siquiera tuvo comienzo.[5]

Específicamente, parece ser que ciertas clases de cosas para las que somos buenos son muy difíciles para los computadores, como el patrón de reconocimiento, la comprensión del habla, recordar y reconocer información relevante y planificar. También ha resultado muy difícil programar computadores para aprender, en cualquier dominio. Por otra parte, los computadores son mucho mejores que los humanos en otras tareas; paradigmáticamente, en el tratamiento de números y, en general, en cualquier cosa que involucre muchas computaciones especificables con precisión o la manipulación de grandes cantidades de datos de acuerdo con reglas precisas.

Esto sugiere que los cerebros humanos no hacen cosas de la manera convencional en la que los computadores lo hacen, que nuestra inteligencia es de una clase diferente. Hay otras dos cosas que los partidarios del conexionismo gustan traer en apoyo de la idea de que la inteligencia humana es de un tipo diferente del de la "inteligencia" computacional: las neuronas son lentas y de degradación armoniosa [*graceful degradation*].

En el mejor de los casos, en tiempos medidos en milisegundos, las neuronas son 10^6 veces más lentas que la presente generación de computadores. El significado exacto de esta comparación no es claro, ya que no tenemos idea de cómo las funciones cognitivas podrían ser implementadas en el cerebro, sin embargo parece cierto que las neuronas

4. "Reglas y representaciones" suena bien, pero es más revelador pensarlo como el punto de vista de las "representaciones y reglas". Hay representaciones y hay reglas que *refieren a esas representaciones mentales.*

5. Cf. Hubert L. Dreyfus y Stuart E. Dreyfus, *Mind over Machine* (Nueva York, The Free Press, 1986).

individuales computan funciones simples. Estamos capacitados para hacer un procesamiento muy sofisticado, tal como la recuperación de memoria [*memory retrieval*], el procesamiento perceptual y el procesamiento lingüístico, en unos pocos cientos de milisegundos. Esto hace que tengamos que aceptar lo que Jerome Feldman[6] llamó la restricción de los 100 pasos para todo programa. Ninguna tarea puede requerir más que 100 pasos seriados, aproximadamente. Esto, al menos, se aplica al nivel cognitivo de las reglas y representaciones, en el que las reglas se refieren a la estructura sintáctica de nuestras supuestas representaciones internas. Incluso, si muchas cosas son hechas en paralelo, esto significa que ninguna de las pistas paralelas puede tener más que 100 pasos, aproximadamente. Si hacemos cosas cognitivas siguiendo programas, ésta es una limitación severa. Probablemente, los conexionistas argumentan, no es así como las hacemos.

"Degradación armoniosa" se refiere a una variedad de maneras según las cuales el funcionamiento [*performance*] de los sistemas naturales se deteriora de manera gradual. Una lesión o una enfermedad cerebral puede afectar el funcionamiento, pero no causar pérdida de recuerdos o de funciones específicos, antes de alcanzar un grado de severidad en el que se pierde una clase entera de capacidades. Y cuando nos aproximamos a los límites de nuestras habilidades en algún área, sea por las restricciones temporales o por la sobrecarga de información, por ejemplo, nuestro comportamiento se ve afectado, pero no se quiebra. El procesamiento de reglas y representaciones, por otro lado, es naturalmente "frágil". Las lesiones causan la pérdida de funciones particulares y de todo lo que dependa de ellas. El sistema trabaja, en un aspecto particular, o no trabaja.

La lentitud de las neuronas, la degradación armoniosa *versus* la fragilidad, y los dominios diferentes en los cuales cada cual está en sus mejores condiciones, sugieren que los sistemas naturales y los computadores convencionales hacen las cosas de maneras muy diferentes. Y los conexionistas han usado estas cosas como argumentos para buscar una aproximación a la cognición que sea mucho más análoga a la manera en que el cerebro trabaja. Pero nada de esto podría haber producido una crisis en BAIA. La crisis tuvo que venir desde dentro.

El problema es que hasta el momento ha resultado imposible encontrar las clases de reglas que BAIA requiere. Las dificultades pare-

6. J. A. Feldman, "Connectionist Models and Their Applications: Introduction", *Cognitive Science* 19, 1-2.

cen generar dos grupos de problemas. El primero es el problema de tratar con restricciones débiles [*soft*] múltiples y simultáneas. El segundo grupo de problemas resulta del hecho de que cualquier porción [*bit*] de conocimiento de sentido común podría resultar relevante para cualquier tarea o cualquier otro conocimiento. Incluidos aquí, quizá como dos extremos de un espectro, están el así llamado problema del "marco" [*"frame" problem*] y lo que llamaré el problema del "cruzamiento"[7] [*"folding" problem*].

Considérese el problema de llegar a una tienda específica en un centro comercial. Uno tiene que trazar un trayecto desde la puerta del centro comercial hasta el negocio, usando quizás escalones o una escalera mecánica, evitando los obstáculos estacionados, como plantas, fuentes, bancos y las áreas abiertas al piso de abajo, y evitando también tropezar con gente, cochecitos para niños y cosas parecidas. Hacemos esto de manera tan natural que es difícil apreciar la complejidad de la tarea. Pero programar un computador para que lo haga requeriría reglas que tomaran en cuenta una enorme cantidad de factores. El número de computaciones requeridas se expande exponencialmente en relación con el número de los factores, conduciendo a una explosión computacional inmanejable.

El ajedrez proporciona un ejemplo bien estudiado de explosión computacional. Los mejores programas de ajedrez disponibles son capaces de examinar alrededor de 10 millones de posiciones cuando buscan cuatro o posiblemente cinco movidas futuras. Para limitar las posibilidades, usan programas ingeniosos de "poda" [*"tree" pruning*], que ignoran movidas implausibles.[8] Pero en el caso del centro comercial, no está del todo claro cómo dar una regla general para ignorar factores o posibilidades.

Típicamente, cuando hay restricciones múltiples y simultáneas, las restricciones son "débiles". Es decir, cualquiera de ellas puede ser violada mientras el sistema cognitivo está haciendo su tarea del modo apro-

7. Estos grupos de problemas están relacionados, y como veremos, en última instancia sus relaciones serán, sin duda, influenciadas por su solución. Ciertamente, poner múltiples restricciones con restricciones débiles ya es ver las cosas desde la perspectiva del conexionismo.

8. Arthur Samuel, autor de un programa de juego de damas, que es uno de los primeros y mejores programas de juego escritos, calculó que si fueran usados métodos de fuerza bruta que tuvieran en cuenta toda movida posible, le llevaría a la computadora más rápida 10 siglos hacer la primera movida. Citado en Dreyfus y Dreyfus, *op. cit.*

piado. El ajedrez, y aun el ejemplo del centro comercial, tal como fueron descriptos brevemente aquí, pueden resultar engañosos al respecto, permítasenos por eso considerar otro ejemplo.

La comprensión del habla involucra muchas clases distinguibles de factores, incluyendo al menos la fonología, la sintaxis, la semántica y varios factores agrupados a menudo en la pragmática, que incluyen al menos el contexto social, el físico y el conversacional. Cada uno de estos "niveles" introduce un sistema complejo de factores que concurren en la interpretación de los estímulos verbales. Pero, en cualquiera de esos niveles, las reglas pueden ser violadas con poco o con ningún deterioro de la comprensión. Por ejemplo, uno puede entender el habla de un vasto número de dialectos regionales, la mayoría de los cuales no podría imitar. Lo mismo vale para la sintaxis y la estratificación social del lenguaje. Lo que uno escucha viola las reglas del lenguaje propio —de cualquier lenguaje que uno podría hablar—, pero uno usa las reglas del lenguaje propio para comprenderlo.[9] Las reglas desempeñan un rol aun cuando sean violadas. Esto es común cuando hay restricciones múltiples.

El segundo grupo de problemas en el cual la BAIA naufraga tiene que ver con el hecho de que cualquier parte del conocimiento de sentido común podría tener que ser usada en relación con cualquier tarea cognitiva. La cognición no puede ser partida en dominios aislados. Esto requiere ser capaz de ver que cierta información es relevante y ser capaz de encontrar información relevante, a partir de la base de datos total que abarca todo el conocimiento que uno posee. Fácil para nosotros, difícil para los computadores.

La IA clásica ha tenido un grado moderado de éxito en escribir programas que "comprenden" relatos [*stories*] simples acerca de dominios específicos.[10] Tienen lo que se llama un "guión" ["*script*"] para ese dominio, que pueden usar para hacer inferencias que les permitan responder preguntas que no están explícitamente respondidas en el relato.

Recientemente, llevé un grupo de niños de seis años al Mundo Marino para una fiesta de cumpleaños. Todos sabían acerca de las fiestas de cumpleaños. Algunos de ellos habían estado en el Mundo Marino y podían describirlo a los otros, pero ninguno había estado en una fiesta

9. Las reglas de la BAIA *son* reglas duras, con condiciones necesarias y suficientes, de estricta aplicación. Hay dispositivos para construir, al menos, una imitación de la debilidad. Pero la debilidad parece ser un rasgo natural de la cognición humana.

10. Roger C. Shank con Peter Childers, *The Cognitive Computer* (Addison Wesley, 1984).

en el Mundo Marino. Eran capaces de combinar sus "guiones" de fiesta de cumpleaños con sus "guiones" del Mundo Marino para planificar las actividades del día, incluyendo cómo compatibilizar las cosas del Mundo Marino con las cosas de una fiesta de cumpleaños, qué cosas típicas de una fiesta serían incluidas, cuáles dejadas a un lado y cuáles modificadas apropiadamente, y qué precauciones se tomarían para mantener al grupo unido ya que después de todo, ellos sólo tenían seis años. Los computadores no pueden hacer nada similar a "cruzar" conocimiento de esta manera, y nadie tiene la clave de cómo lograr que lo hagan.

En el otro extremo del espectro, cualquier porción aislada de información que adquiramos cambiará nuestro sistema total de creencias. Algunas de las creencias más tempranas tienen que ser borradas y tendremos que hacer algunas inferencias obvias a partir de otras creencias. Pero la nueva porción de información será irrelevante para la mayoría de nuestras creencias previas, y éstas permanecerán sin cambios. El problema del marco es el problema de determinar, de una manera efectiva y general, qué cambiaría cada nueva porción de información y qué dejaría igual en un sistema de creencias.[11] Y la primera parte del problema es la de determinar cuál de las viejas creencias son de alguna manera relevantes para la nueva información. Para un sistema de cualquier dimensión, es totalmente ineficiente el tener que buscar en cada porción de vieja información para ver si está afectado y cómo lo está.

Porque combinamos el conocimiento de manera tan natural y las inferencias parecen tan naturales y obvias, puede resultar difícil ver el problema si uno no ha tratado de abordarlo. El punto puede plantearse de esta manera. Considérese cualquier programa que se proponga imitar el comportamiento de algún ser humano normal. Si no accede a todo el conocimiento de esa persona, será pasible de contraejemplos. Además, aun si puede acceder a todo ese conocimiento, tiene también que especificar, de algún modo, *por adelantado* qué cambiaría cada nueva

11. La nueva información también puede cambiar cómo uno representa una tarea. Ver un animal dañino que uno quiere evitar cambia el problema del trazado del trayecto en el centro comercial. Tener un encuentro casual con una persona conocida, no. Por supuesto, podríamos construir esto en nuestro problema de circulación en el centro comercial. El problema es que, virtualmente, cualquier información nueva podría, dados otros factores, cambiar la tarea entre manos. Parece que nuestro programa de circulación en el centro comercial va a tener que tener reglas que determinen, para cualquier situación específica posible, si cada porción particular de nueva información potencial es o no relevante y si esto es así, a qué cambios induce.

porción de conocimiento y qué dejaría igual, en cualquier situación. Sin eso, sería pasible, nuevamente, de contraejemplos sistemáticos.

Lo que puede ser el lado resbaladizo [*flip*] de este problema es el hecho de que virtualmente no hay límites sobre qué parte del conocimiento especializado o de sentido común podría ser relevante para resolver un problema particular. Somos tan buenos en recuperar y usar información relevante que notamos los fracasos y tomamos el éxito como una rutina. Por otro lado, esto es muy difícil para los computadores convencionales. La recuperación de memoria está en algún sentido basada en el contenido. En la arquitectura de un computador convencional, la recuperación se hace basada en la dirección [*address*] en donde el dato está almacenado. Así, parte del problema para BAIA es el problema de la memoria direccionable al contenido [*content addressable memory*].

Pero esto es sólo parte del problema y darle ese rótulo puede oscurecer la dificultad. Primero, uno tiene que ver que una porción de información es relevante. Entonces la recupera. Pero, ¿cómo puede ver que una porción de información es relevante hasta que la ha encontrado? Aparentemente uno tiene que encontrarla y ver que es relevante al mismo tiempo. Como dijo Hume, "Uno pensaría que todo el mundo intelectual de las ideas fue expuesto al mismo tiempo a nuestra vista y que no hicimos sino escoger aquello que fue lo más apropiado para nuestros propósitos" (*Tratado* I, I, VII). Lo que necesitamos es memoria direccionable a la *relevancia*, sea eso lo que fuere.

También, a veces, todo este grupo de problemas y el problema del marco, en particular, es caracterizado como el problema de representar el conocimiento de sentido común. Pero esto, también, puede minimizar las dificultades. Porque es el problema de representar, almacenar, recuperar y usar el conocimiento, todo en un mismo paquete.

Estos dos grupos de problemas, el de las restricciones débiles, múltiples y simultáneas y el de la relevancia potencial de todo con todo, no *prueban* que los supuestos de la BAIA estén equivocados, y que las reglas y representaciones no puedan proveer un modelo satisfactorio de la cognición humana. El punto es que tales problemas han ofrecido resistencia al progreso serio por más de una década, y la atracción del conexionismo debe ser entendida bajo esa luz.

3. *El ABC del conexionismo*[12]

Un sistema conexionista, o una *red neuronal* consiste en una red de procesadores simples similares a neuronas [*neuron-like processors*], llamados *nodos* o *unidades*. Cada nodo tiene conexiones dirigidas a varios otros nodos, de modo que obtiene señales de algunos nodos y envía señales a otros nodos, incluyendo, posiblemente, aquellos de los cuales obtiene señales. En la práctica, un nodo dado puede obtener *input* de sólo dos o tres nodos, o de tantos como dos o tres docenas. En principio, podrían ser miles. El *input* de cada nodo es una señal simple como una corriente eléctrica o una transmisión sináptica. Esta señal podría tener sólo los valores encendido y apagado [*on and off*] o podría variar en una sola dimensión, la cual llamaré *fuerza* [*strength*]. El *input* total de un nodo determinará su *estado de activación* [*state of activation*]. El nodo puede estar encendido o apagado. Podría estar encendido toda vez que obtenga un *input* cualquiera, o podría haber un umbral que el *input* tiene que alcanzar antes de que el nodo se encienda. En cualquiera de estos casos el nodo podría tener sólo dos valores, "encendido" y "apagado", o el estado de activación podría variar como una función del *input* total. Finalmente, algunas veces los sistemas admiten señales que inhiben la activación así como el *input* excitatorio.

Cuando un nodo está encendido, manda señales a los nodos con los que tiene conexiones de *output*. La fuerza de la señal de *output* es una función de su grado de activación y, nuevamente, varias funciones han sido usadas.

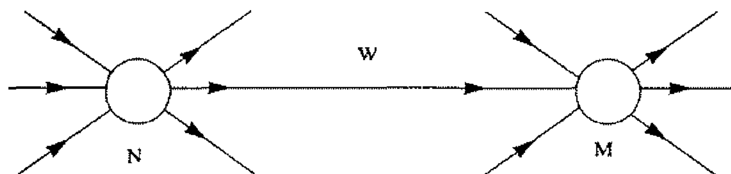


Figura 1. Hay una conexión simple, unidireccional, entre los nodos, que está indicada aquí por las flechas. El nodo N recibe *input* de muchos nodos. Su señal de *output* a lo largo de cada línea de *output* es una función de su *input* total. El *input* al nodo M desde el nodo N es una función del *output* de N sobre la conexión N-M y el peso W de esa conexión.

12. La biblia del conexionismo es David E. Rumelhart, James L.McClelland y el Grupo de Investigación PDP, *Parallel Distributed Processing*, 2 Volúmenes (Cambridge, Bradford Books, MIT Press, 1986), en adelante PDP. Los primeros cuatro capítulos de PDP, Parte I (págs. 3-146), introducen "The PDP Perspective".

Las conexiones entre los nodos son análogas a las conexiones de los cables eléctricos o sinápticos. Así, tienen un grado de resistencia. El *input* al nodo *b* desde el nodo *a* es una función de la fuerza de la señal de *output* de *a* y la fuerza de la conexión entre ellos. La fuerza de la conexión entre los nodos es referida como *peso* [*weight*]. Un peso mayor significa menos resistencia: una señal más fuerte recibida.

Típicamente, en un sistema conexionista dado, las propiedades de los nodos son consideradas fijas. Los pesos, por otro lado, pueden variar como una función de la experiencia. Como resultado, es posible construir sistemas conexionistas que sean capaces de aprender de maneras muy naturales. De este modo, el aprendizaje conexionista consiste en "obtener que nuestros pesos cambien".

Para dar un ejemplo simple, si el peso entre dos nodos aumenta, significando que la resistencia es menor, siempre que ambos nodos estén encendidos juntos, aumenta la probabilidad de que cuando uno esté encendido el otro también se encienda. Esto vale para el aprendizaje simple de tipo asociacionista.[13] En un momento discutiremos algunos otros tipos de aprendizaje.

En un sistema conexionista, algunas unidades obtienen estímulos desde fuera del sistema. Éstas son unidades de *input*. Algunas son unidades de *output*. Se las piensa como enviando señales fuera del sistema. Las unidades de *input* pueden también obtener estímulos de nodos de dentro del sistema y los nodos de *output* pueden enviar una retroalimentación a nodos del sistema. El resto de los nodos, sin conexiones fuera del sistema, son llamados nodos *ocultos* o *internos* [*hidden or internal nodes*].

El marco [*framework*] conexionista no impone limitaciones sobre los tipos de las estructuras de nodos que son posibles, y en los volúmenes PDP pueden ser encontradas muchas clases diferentes de estructuras. Las estructuras por capas [*layered structures*], como la de la figura 2, son útiles como ejemplo. En tal estructura, se plantea un problema mediante la activación de un patrón de nodos de *input*, el procesamiento procede a través de la activación de nodos en una o muchas, típicamente, más pequeñas capas ocultas hasta que el patrón de activación aparece en los nodos de *output*. Esto cuenta como la solución que da el sistema al problema planteado. Las complicaciones posibles incluyen la retroalimentación, lazos [*loops*] y las conexiones inhibitorias dentro de

13. [...] George Graham sugiere que el conexionismo puede proporcionar un entendimiento más profundo de, al menos, algunos aprendizajes simples, que el proporcionado por el conductismo y la teoría de la discrepancia.

un conjunto de unidades, de suerte que sólo una o un grupo fijo de unidades pueden ser encendidas juntas: "el que gana se lleva todo".

En un procesamiento típico no se da el caso de que sólo sean activados los nodos internos y de *output* correctos. Típicamente, hay mucha actividad, con nodos encendiéndose y apagándose, enviando y recibiendo señales repetidamente hasta que el sistema "se asienta" en una configuración estable que constituye su solución al problema planteado. Imagínese un grupo de luces conectadas, encendiéndose y apagándose hasta alcanzar gradualmente un estado estable con algunas encendiéndose y algunas apagándose.

Varias diferencias entre el conexionismo y la arquitectura clásica son aparentes. En la arquitectura clásica, lo que ocurre en todo el sistema es controlado por el programa que está en la unidad de procesamiento central (UPC). En un sistema conexionista, no hay un ejecutor central, y no hay un programa que determine lo que ocurre en el sistema. Todas las conexiones son locales, de modo que cada nodo sólo sabe lo que obtiene de los nodos con los cuales está conectado. Y la única información que un nodo comunica a otro es "Estoy encendido, (tanto)". Nadie en el sistema sabe más que eso. Así, en particular, nadie en el sistema sabe lo que el sistema como una totalidad, está haciendo. Además, lo que ocurre en un lugar del sistema es independiente de lo que ocurre en otra parte del mismo. El comportamiento de cada nodo está determinado sólo por su estado actual y su *input*. Sin embargo, los sistemas conexionistas pueden ser interpretados como representando globalmente contenido interesante cuando se encuentran en un estado particular, y como teniendo almacenado (en los pesos) conocimiento que no está representado en ese momento de manera activa.

"Conexionismo" es un buen nombre para la familia de sistemas que se ajustan a esta descripción. Por otro lado, aunque igualmente aceptado, el término "paralelo", tal como aparece en "masivamente paralelo", o aun en "Procesamiento Distribuido en Paralelo", me parece menos afortunado. Puede llevar a confusiones dado que puede sugerir muchas *secuencias* separadas que van en paralelo, cada una de acuerdo con su propio programa. Ésta no es la descripción correcta de los sistemas conexionistas. Ellos no involucran procesos paralelos independientes que estén programados o bien ínsitos aunque en principio programables.[14] Se caracterizan por un procesamiento local simultáneo distribui-

14. El procesamiento de nodos y los cambios de pesos en un sistema conexionista son ambos algorítmicos. Así, por supuesto, el procesamiento conexionista es programable.

do sobre todo el sistema. Desde este procesamiento local simultáneo emerge una configuración estable del sistema, la cual constituye la solución del sistema al problema planteado.

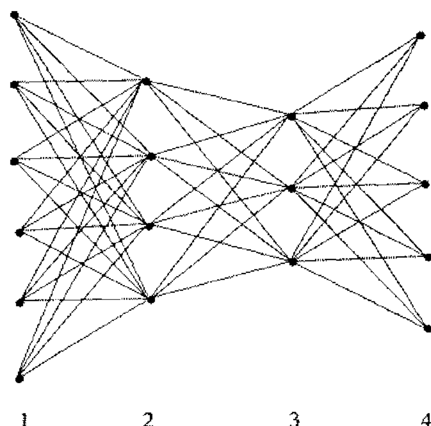


Figura 2. Una red de alimentación progresiva [*feed forward network*]. Los nodos de la capa 1 son nodos de *input*. Los nodos de la capa 4 son de *output*. Cada nodo en la capa n tiene conexiones de *output* con cada nodo de la capa n+1. No hay otras conexiones en la red. Cuando los nodos en la capa 1 reciben el *input*, la activación se desparrama desde la capa 1 a las capas 2, 3 y 4.

Las redes de alimentación progresiva, como la de la figura 2, son capaces de aprender de una manera muy natural, y han merecido un amplio estudio.[15] A los nodos de *input* y *output* les son asignadas arbitrariamente interpretaciones, de tal modo que ciertos patrones de nodos de *output* serán el *output* "correcto", dados ciertos patrones de estimu-

---

Además, los sistemas conexionistas pueden ser simulados en computadoras convencionales y ciertamente, en este punto, la mayoría de las investigaciones conexionistas siguen esta línea. Así, hay, como tiene que haber, un nivel de análisis en el cual los sistemas conexionistas son controlados (o controlables) por reglas o programas. Sin embargo, la observación que hago en el texto es que esto no es lo que sucede en un sistema conexionista, en el nivel del procesamiento cognitivo de las representaciones. Todo el procesamiento es local, determinado sólo por nodos y pesos vecinos. La solución del sistema a un problema, *emerge* de ese procesamiento local...

15. Cf. "Learning Internal Representations by Error Propagation", D.E. Rumelhart, G.E. Hinton y R.J. Williams, capítulo 8 de PDP.

lación de los nodos de *input*. Los pesos son fijados inicialmente en valores pequeños tomados al azar, a través de todo el sistema. En este punto, será puramente accidental si el sistema produce el *output* correcto para un *input* dado. Ahora bien, los pesos son ajustados algorítmicamente capa por capa a través del sistema, como una función de la contribución de cada peso al error. Para una amplia gama de problemas, después de pasar repetidamente por este procedimiento de "retropropagación" ["*back propagation*"], el sistema producirá el *output* correcto para todo *input*. Habrá "aprendido" las respuestas y puede aun ser capaz de "generalizar" a casos que no ha visto.[16]

El éxito de la retropropagación como una técnica de entrenamiento muestra que los sistemas conexionistas son capaces de un aprendizaje interesante. Pero la retropropagación, tal como se la entendió hasta ahora, es insatisfactoria en varios aspectos. Si la red es concebida como un modelo de un sistema natural, tiene que haber algún proceso *en la red* que cambie los pesos apropiadamente. En las simulaciones actuales no se provee ningún medio tal, y no está claro cómo esos procesos podrían ser construidos dentro de una red. La retropropagación es también extremadamente lenta, pues requiere una gran cantidad de pasos para un aprendizaje exitoso. Y queda aun algo de magia negra. Lo que funciona y lo que no funciona, tanto en términos de una estructura de red como en términos de parámetros matemáticos, depende en gran medida de la intuición del experimentador y del ensayo/error. Sin duda, todo esto se debe, al menos en parte, a la juventud de la empresa y constituye un área activa de la investigación conexionista.

4. La representación en los sistemas conexionistas

Geoffrey Hinton ha descripto un sistema interesante que aprende las relaciones de parentesco familiar por retropropagación.[17] El resto de

16. Cf. "On Learning Past Tenses of English Verbs", D.E. Rumelhart y J.L. McClelland, capítulo 18 de PDP, volumen 2. Aunque artificial en varios aspectos, el sistema descripto en ese artículo aprende en una secuencia interesante como aquella en la cual los niños aprenden realmente los tiempos pasados. Primero, obtiene de manera correcta formas irregulares comunes. Después, patrones regulares "sobregeneralizados" y produce formas como "*goed*".

17. Cf. por ejemplo, "Learning Distributed Representations of Concepts", *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, Amherst, MA, 1986, págs. 1-12; y "Learning in Parallel Networks", *Byte*, abril de 1985, págs. 265-273.

los conceptos conexionistas que se introducirán, serán discutidos en términos de ese sistema. Los datos que el sistema aprende consisten en dos árboles de familia isomórficos de tres generaciones, con 12 individuos en cada familia.

La red tiene 5 capas. La capa de *input* consiste en dos grupos distintos, uno con 24 nodos, asignados a uno de los 24 individuos, el otro grupo tiene 12 unidades, asignadas a las relaciones padre, madre, esposo, esposa, hijo, hija, tío, tía, hermano, hermana, sobrino y sobrina. Hay 24 nodos de *output*, uno por cada individuo. Las tres capas internas u ocultas consisten de 12, 12 y 6 unidades respectivamente. Salvo una excepción, cada capa está totalmente conectada con la capa subsiguiente, y no hay otras conexiones en la red. La excepción es que la primera capa interna, esto es, la segunda capa está dividida en dos grupos de seis unidades, un grupo conectado sólo con unidades de relaciones, el otro grupo totalmente conectado con unidades de personas.[18]

El *input* siempre está dado por el encendido de un nodo de persona y de un nodo de relación, digamos, "Penélope" y "esposo". El sistema obtiene la respuesta correcta, si el único nodo de *output* que se enciende es el nodo para su esposo, esto es, "Cristóbal". Si el *input* fuera "Colin" y "tía", la respuesta correcta constaría de dos nodos de *output* específicos, ya que Colin tiene dos tías.

Inicialmente, los nodos internos no tienen contenido representacional. Los pesos son establecidos con valores de un grado mínimo de azar. Al principio, por supuesto, el *output* es como conjeturar: la mayoría de las respuestas son equivocadas. Los pesos son lentamente cambiados en la dirección de las respuestas correctas.[19]

El sistema fue "entrenado" en 100 de las 104 relaciones codificadas en los árboles de las dos familias. En cierto momento, después de 1500 pasadas de los 100 casos, el sistema adquirió correctamente todas esas relaciones. Además, fue capaz de generalizar o de hacer la

18. Conectadas totalmente significa que toda unidad en la capa n está conectada con toda unidad en la capa n+1. No hay una razón a priori para esta red particular. Elegir una estructura es una cuestión de intuición del investigador y de ensayo/error. Un principio es que para inducir a la generalización las capas internas tienen que contener *menos* nodos que las capas de *input* y *output*.

19. Tal como se hizo en realidad, corrieron en total cien casos antes de que cualquier peso fuera cambiado. Entonces, cada peso fue cambiado como una función de su contribución al error total. Los detalles de ese cálculo no interesan aquí, aunque tales asuntos son el centro de gran parte de las investigaciones conexionistas. Ver Hinton, *op. cit.*

inferencia apropiada, y obtener con corrección las cuatro relaciones restantes.[20]

Esto puede no parecer demasiado impresionante como una "generalización", pero es aprendizaje y uno no debería esperar mucho más de este caso particular. Lo que ha sucedido es esto. Para obtener de manera correcta los 100 casos en los que fue entrenado, con sus recursos neurales limitados, el sistema tuvo que codificar la estructura del sistema de parentesco y los rasgos relevantes de los individuos. Habiendo hecho esto, fue capaz de "inferir" las relaciones que no se le habían comunicado. Es decir, la información que había codificado determinó esas relaciones. Dado que comenzó como una *tabula rasa* se le dieron suficientes datos como para determinar la estructura del dominio con el que estaba tratando.

Después del entrenamiento, muchos nodos internos tuvieron un contenido discernible, por ejemplo, "primera generación" o "inglés", o "rama izquierda del árbol de familia".[21] Es decir, hubo un único nodo interno que sólo se encendería cuando la persona dada como *input* fuera inglés, etcétera. (Además, el sistema aprovechó el hecho, que por supuesto podría no habérsele comunicado, de que los árboles de las dos familias eran isomórficos.)

A los nodos de *input* y *output* se les asigna arbitrariamente ciertos significados. *Dada* esa interpretación, se puede decir que el sistema ha adquirido los conceptos representados por sus nodos internos. También podemos decir con el mismo engreimiento, que el sistema aprendió que Penélope es inglesa, de primera generación, etcétera. Estas revelaciones se producen cuando se le da "Penélope" como *input*. La "reconoce" como inglesa.

No todos los nodos internos del sistema final tienen un contenido obvio, describible. Pero en este pequeño sistema, todos los nodos internos son necesarios para que el sistema funcione apropiadamente. Si se hubiese suprimido alguno de los nodos de una red entrenada, ella no hubiera obtenido todas las respuestas correctas. Por lo tanto, parecería que algunos nodos estaban respondiendo a rasgos reales del dominio, aunque no fácilmente describibles.[22]

20. En un ensayo dio la respuesta correcta en tres de los cuatro casos en los cuales no fue entrenada. En el otro, dio respuesta correcta a todos los cuatro casos de prueba.

21. Un punto interesante es que ninguno de los nodos de la segunda capa representaron "femenino", probablemente esto fue porque todas las relaciones usadas estaban determinadas por sexo.

22. La habilidad de hacer esto es ciertamente una potencialidad atractiva del conexionismo...

Otro hecho interesante acerca de esta red es que cuando está entrenada más de una vez partiendo de pesos iniciales diferentes, puede concluir con representaciones internas diferentes. Concluirá, ciertamente, con pesos finales diferentes. Pero también ciertos conceptos pueden ser codificados por nodos internos diferentes y de manera más interesante aun, puede no haber en un sistema entrenado, un análogo de un concepto codificado en otro. Por eso, las mismas relaciones de *input-output* pueden ser logradas usando conceptos internos diferentes.[23] Hay diferentes maneras de codificar la estructura del dominio. Es razonable sospechar que esto es un rasgo común de las redes conexionistas y de los sistemas naturales.

Pasaré ahora a un último rasgo, muy discutido, del pensamiento conexionista: la "codificación gruesa" ["*coarse coding*"]. Podemos pensar nuestra red como teniendo una representación *interna* de Penélope. Un cierto patrón de activación, con algunos nodos encendidos y algunos apagados, es propio de Penélope. Ella es la única que es inglesa, de primera generación, etcétera. Algunos de los nodos que están encendidos para Penélope lo están también para Andrés. Pero Andrés tiene un patrón de activación total diferente. Algunos nodos que están encendidos para Penélope están apagados para él, y otros nuevos están encendidos. Ciertamente, hay un patrón de activación interno propio para cada individuo.

A las representaciones que involucran muchas unidades se las llama representaciones *distribuidas*. Cuando el mismo nodo es parte de muchas y diferentes representaciones distribuidas, como en nuestra red, tenemos lo que se denomina *codificación gruesa*.[24] Con la codificación gruesa, los nodos pueden o no tener contenido de manera individual. En nuestra red, hay también representaciones *locales*, nodos singulares, de los individuos. En las redes con representaciones distribuidas es más común, sin embargo, que los contenidos sean representados sólo de una manera distribuida.

Las representaciones distribuidas suscitan muchas posibilidades interesantes, de las cuales sólo mencionaré un par. Si una representación

23. Uno recuerda aquí la forma mastodóntica de la que habla Quine, que satisfacen diferentes arbustos. *Word and Object* (Cambridge, MIT Press, 1960), pág. 8.

24. Éste me parece el uso más razonable y útil de este término, pero no es universal. Algunas veces se piensa a los nodos individuales como teniendo contenido específico, pero en el nivel de conceptualización: micro-rasgos. Entonces, los nodos compartidos de las representaciones distribuidas pueden ser pensados como representando contenido compartido o sobrelapado.

dada es codificada como la actividad de un gran número de nodos, puede que no sea necesario para todos esos nodos estar activos para constituir la activación de la representación. Para tomar el caso más simple, la activación de la representación podría consistir en la activación de un número suficiente de sus nodos. Así, el mismo contenido representacional puede estar presente en un sistema como la actividad de algunos nodos diferentes en ocasiones diversas.

Un rasgo de muchos sistemas con representaciones distribuidas es que los pesos son establecidos de modo tal que las representaciones tienen una tendencia a completarse a sí mismas, una vez que algunos de sus nodos están activos. Esto coadyuva para algo así como la memoria direccionable al contenido. Una porción del contenido de una representación compleja tiende a activar la totalidad.

Cuando los nodos que son parte de varias representaciones complejas son activados, puede darse entre esas representaciones una competencia por la consumación [*completion*]. Ésta es una de las situaciones en las que un caso de un ganador toma todo el mecanismo que pueda ser operativo. Es también una situación típica en la que hay un proceso de "asentamiento". El sistema chispeará y chisporroteará antes de que se "asiente en" la representación que *mejor se ajusta* al *input*. Puede haber muchos factores diferentes, restricciones, que van determinando cuáles representaciones se ajustan mejor.

Además, algunas de esas restricciones podrían ser violadas por la representación que mejor se ajusta a ellas como un todo. Por ejemplo, la representación interna que gana como la mejor representación del *input* puede, sin embargo, entrar en conflicto con alguna otra del *input*. Es decir, es natural en los sistemas conexionistas que las restricciones sean *débiles*; que es violable cuando el sistema está haciendo su trabajo de manera apropiada. En BAIA las mismas restricciones serían representadas más naturalmente como inconsistencias lógicas y en consecuencia inviolables (cf. nota 9). Los conexionistas creen, pienso que correctamente, que la mayor parte de la cognición es débil. Los rasgos correlativos de las restricciones débiles y que mejor se ajustan, son un motivo importante de atracción del conexionismo.

Supóngase, por ejemplo, que tenemos un sistema conexionista de reconocimiento del habla.[25] Uno esperaría que tal sistema entienda

25. No intento sugerir que ésta sea una posibilidad a corto plazo. Un sistema realista de reconocimiento del habla, conexionista o de otro tipo, puede estar a siglos de distancia o más allá del alcance de los humanos, no obstante los frecuentes pronunciamientos optimistas de la comunidad de BAIA. Por todo lo que sabemos ahora, puede ser que la única

acentos nuevos, dialectos y alteraciones del habla novedosos de manera natural y automática, usando los mismos principios que usa en la comprensión del habla que es normal para él. Uno esperaría que un sistema de reconocimiento del habla de BAIA fuera capaz de hacer esas cosas sólo con una programación especial para cada nueva variación, lo cual es hacer trampa.

Podríamos decir, no sólo que el sistema de Hinton ha adquirido el concepto "inglés", sino que *cree* que Penélope es inglesa. La luz "inglesa" se enciende cuando la luz "Penélope" se enciende. También sabe que su esposo es Cristóbal.

¿Dónde está la información de que Penélope es inglesa o que Cristóbal es su esposo, cuando esa información no está activamente representada? En un sentido, no está en ningún lado. En otro sentido, está en los pesos.

En una arquitectura computacional convencional, la información está almacenada de la misma forma en la cual es usada. Cuando se está usando información, una cierta estructura, un objeto sintáctico, está en la UPC. Cuando no se la está usando, la misma estructura de datos está almacenada en la memoria.

En un sistema conexionista, la información está activamente representada como un patrón de activación. Cuando esa información no está en uso, ese patrón no está presente en ningún lugar del sistema. La información no está almacenada como estructuras de datos. Los únicos símbolos siempre presentes en un sistema conexionista son las representaciones activas.

Pero, por supuesto, hay un sentido según el cual la información está en el sistema. Los pesos, la fuerza de las conexiones entre los nodos, son tales que las representaciones activas apropiadas son creadas cuando se las necesita. Así, en nuestro simple ejemplo, cuando el nodo Penélope se enciende, el sistema activa sus representaciones internas de Penélope.

Hay aquí dos cuestiones. La información está almacenada en los pesos. Y porque los pesos son como son, las representaciones son creadas en respuesta a estímulos internos o externos. Los recuerdos no están almacenados, son recreados una y otra vez en respuesta a aquello que nos lo haga recordar. La información que no está efectivamente activa, no en uso, está en el sistema sólo de manera potencial.[26] Por

---

manera de que un humano pueda hacer un sistema de reconocimiento del habla es por copulación.

26. Cuando leí primero esta idea en Locke (*Essay*, II, X) pensé que era zonza. No parece tan zonza ahora.

ende, no hay una distinción interna al sistema entre recrear viejas representaciones y crear representaciones nuevas en respuesta a una situación.

Cualquiera sea el caso respecto de la memoria, esa es seguramente la manera en que las cosas funcionan en muchos dominios cognitivos, como en la percepción y la comprensión del lenguaje. No podríamos almacenar información lingüística acerca de cada oración individual que podamos comprender. En la comprensión de una oración, creamos información lingüística sobre la marcha, en respuesta a los estímulos presentes. Excepto en el caso de unos pocos clichés, no interesa si hemos oído antes la oración o no.[27]

He dicho que el sistema entrenado de Hinton "cree" que Penélope es inglesa. Pero esta manera de hablar es excesiva en una medida en que no lo es respecto de la arquitectura clásica. Cuando la luz "Penélope" se enciende, la luz "inglesa" se enciende. Se podría pensar que esta secuencia temporal es como una oración *hablada*, ya que los computadores convencionales tienen estructuras de datos que son análogas a las oraciones escritas. Pero esto es incorrecto. La secuencia de nodos encendida "Penélope", "inglesa", es *sólo* una secuencia causal. No hay aquí ninguna estructura sintáctica.

Esto señala un rasgo típico de los sistemas conexionistas existentes: las representaciones no tienen, típicamente, ninguna estructura sintáctica. Las representaciones son o bien nodos aislados (codificación local) o bien conjuntos de nodos (codificación gruesa). Cuando se necesita una estructura, usualmente ésta se construye dentro del sistema con diferentes grupos de nodos dedicados a partes diferentes de la estructura que se necesita. Así, en un sistema de reconocimiento de palabras, por ejemplo, habrá un grupo de nodos para la primera letra, un grupo para la segunda, etcétera.[28] Lo que se logra de esta manera es una estructura de representaciones atómicas, no de representaciones estructuradas.

27. Los partidarios de BAIA saben esto, por supuesto. Sin embargo, puede ser significativo que esa sea la manera natural para los sistemas conexionistas.

28. Cf. por ejemplo, J.L. McClelland y D.E. Rumelhart, "An Interactive Activation Model of Contex Effects in Letter Perception: Part I, An Account of Basic Findings", *Psychological Review* 88 (1981). El punto que se hace aquí vale respecto del artículo de Touretzky y Hinton, frecuentemente citado, "Symbols among the Neurons: Details of a Connectionist Inference Architecture", *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 1985.

La falta de representaciones estructuradas sintácticamente no es una limitación esencial[29] de la arquitectura conexionista. Ciertamente, Paul Smolensky ha mostrado una manera elegante de hacerlo, llamada "representaciones de producto vectorial"[30] ["*vector product representations*"]...

Sin embargo, Fodor y Pylyshyn[31] argumentaron que si uno construye representaciones estructuradas sintácticamente en una red conexionista, estará implementando, meramente, la IA clásica (BAIA) de una manera novedosa. Los procesos que dependen de esa estructura serán describibles en términos de secuencias de reglas que corresponden a un programa clásico. Pero entonces, esas reglas o ese programa describirán los procesos *cognitivos* en cuestión. Así desarrollado, el conexionismo no sería una contribución a nuestra comprensión de la cognición.[32]

Supóngase que tuvimos una teoría BAIA exitosa de la cognición humana y una implementación conexionista exitosa que pensamos que revelaba cómo la teoría BAIA podría ser realizada en cerebros humanos. Si encontráramos marcianos que satisficieran la teoría cognitiva BAIA pero no la implementación conexionista, diríamos (y deberíamos decir) que los marcianos eran cognitiva y psicológicamente como nosotros...

Fodor y Pylyshyn argumentan de manera enfática que los procesos cognitivos dependen de representaciones estructuradas. Así, el conexionismo parecería enfrentar un dilema potencialmente devastador. O nos da sólo asociaciones de representaciones no estructuradas o nos da procesos cognitivos que involucran representaciones complejas. Si es lo primero, el conexionismo es un "mero asociacionismo", demasiado débil

29. Algunos, supongo, no verán esto como una limitación. Creo que lo es...

30. Paul Smolensky, "On Variable Binding and Representation of Symbolic Structures in Connectionist Systems" y "A Method for Connectionist Variable Binding", Institute of Cognitive Science, University of Colorado, Technical Report number CU-CS-356-87 y CU-CS-355-87.

31. Jerry Fodor y Zenon Pylyshyn, "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition* 28 (1988), págs. 3-71.

32. Si esto es correcto, la implementación conexionista podría ser aun interesante e importante. Dado que los sistemas conexionistas parecen ser, desde una perspectiva computacional, más parecidos al cerebro que los sistemas de BAIA, implementar sistemas BAIA en redes conexionistas podría ser un paso hacia la comprensión de cómo los procesos cognitivos de BAIA podrían realizarse en el cerebro. Pero eso no sería un paso hacia la comprensión de los procesos cognitivos en sí mismos.

para darnos todo o aun mucho de la cognición.[33] Si es lo último, es una "mera implementación", no informativa en el nivel de la cognición.

Si el conexionismo no es un modelo de toda la cognición humana, entonces, ¿de cuánto de ella lo es? David Kirsch ha señalado que la implementación de representaciones complejas es, desde el punto de vista computacional, enormemente costosa para el conexionismo. Por ende, sugiere que optar a favor del conexionismo equivale a adoptar la hipótesis de que sólo una pequeña porción de la cognición involucra reglas y representaciones estructuradas y que la cognición es en general débil y asociativa.

5. Conexionismo. Abril. 1988

Hay ciertos rasgos deseables que uno esperaría que tengan los sistemas conexionistas, que de hecho son exhibidos por algunos de los sistemas conexionistas actuales: habilidad para aprender, habilidad para tratar con múltiples restricciones débiles, patrones de reconocimiento exitosos y memoria direccionable al contenido. Éstos son algunos de los factores que parecen estar en el corazón de la crisis de BAIA.

A mi entender, ningún sistema conexionista producido hasta ahora debería ser interpretado como un intento serio para modelar o simular cualquier aspecto de la cognición humana. Los sistemas conexionistas que he visto son demasiado simples, con demasiadas restricciones artificiales y han recibido demasiada ayuda por una especificación artificial del dominio de la tarea. Además, no hay perspectiva de que esos sistemas "alcancen" a modelar algún aspecto de la cognición real.

Sin embargo, esos sistemas muestran que los sistemas conexionistas pueden hacer al menos algunas de las clases de cosas correctas. Además, algunos de estos sistemas exhiben propiedades naturales intrigantes.[34] Así, sirven como argumentos plausibles. Muestran que vale la pena pensar seriamente acerca de cómo el conexionismo tendría que ser desarrollado para dar un modelo serio de (toda o de cualquier parte de) la cognición.

Desde este punto de vista, debemos tener presente que el conexio-

33. Los conexionistas sostienen que están presentando un modelo de la cognición en general, por ejemplo, PDP, capítulo 4, pág. 110.

34. Cf. por ejemplo, el sistema de aprendizaje de los tiempos pasados en Rumelhart y McClelland en PDP, II, capítulo 18.

nismo luce atractivo en parte, precisamente, porque luce prometedor allí donde BAIA fracasó repetidamente. Si el conexionismo ha de tener éxito como un modelo de la cognición, tiene que plantearse los problemas que han precipitado la crisis en BAIA. Si el conexionismo no puede resolver esos problemas, no merece reemplazar a BAIA.

Quizá, se pueda decir que algún progreso se hizo en el área de los problemas de reconocimiento. Sin embargo, ningún progreso se hizo en los problemas del marco y del cruzamiento. Ciertamente, debido a que estos problemas involucran información proposicional es decir, estructurada sintácticamente, no pueden ni siquiera ser formulados en los términos de los sistemas conexionistas típicos producidos hasta ahora.[35*]

TRADUCTORA: Liza Skidelsky.
REVISIÓN TÉCNICA: Eduardo Rabossi.

35. Esto se aplica a los sistemas exhibidos, en contraste con la discusión intuitiva, en el lugar en donde uno más esperaría alguna claridad sobre este asunto. "Schemata and Sequential Thought Processes in PDP Models", por Rumelhart, Smolensky, McClelland y Hinton, PDP, II, capítulo 14.

* Deseo agradecer a Stan Franklin por los comentarios hechos a un borrador anterior a este artículo, que condujeron a varias mejoras.

Capítulo 14

LA ESTRUCTURA CONSTITUTIVA DE LOS ESTADOS MENTALES CONEXIONISTAS: UNA RESPUESTA A FODOR Y PYLYSHYN *

Paul Smolensky

El principal propósito de este artículo es responder al punto central de la crítica de Fodor y Pylyshyn (1988) al conexionismo. La respuesta directa a sus críticas comprende la Sección 2 de este trabajo. En suma, argumento que Fodor y Pylyshyn simplemente están equivocados en su alegación de que los estados mentales conexionistas carecen de la necesaria estructura constitutiva [*constituent structure*], y que el origen de ese error está en no apreciar la significación de las representaciones distribuidas [*distributed representation*] en los modelos conexionistas. La Sección 3 es una respuesta más amplia al punto básico de sus críticas, que consiste en que los conexionistas deberían reorientar sus trabajos hacia la *implementación* de la arquitectura cognitiva simbólica clásica. Argumento, en cambio, que la investigación conexionista debería desarrollar *nuevas formalizaciones* de las nociones computacionales fundamentales, a las que se les ha dado un molde formal particular en el paradigma simbólico tradicional.

Mi respuesta a la crítica de Fodor y Pylyshyn supone un cierto contexto meta-teórico que es expuesto en la Sección 1. En esta primera sección argumento que cualquier discusión acerca de la elección de un marco teórico [*framework*] para la modelación cognitiva (por ejemplo, el marco conexionista) tiene que admitir que tal elección involucra una respuesta a una paradoja cognitiva fundamental y que esa respuesta da forma a la empresa científica que rodea a la investigación en ese marco teórico. Fodor y Pylyshyn favorecen, implícitamente, una clase de respuesta a la paradoja, en oposición a otra, y deseo analizar sus críticas bajo esa luz.

* "The Constituent Structure of Connectionist Mental States: A Replay to Fodor and Pylyshyn", *The Southern Journal of Philosophy*, 26 (supl.), (1987), págs. 137-161. Con autorización del autor y del *Southern Journal of Philosophy*.

1. La paradoja y varias respuestas

En esta sección, quiero considerar la pregunta acerca de qué factores entran en la decisión relativa al formalismo modelador cognitivo a adoptar, dada la elección entre el formalismo simbólico y el formalismo conexionista. Quiero argumentar que la movida crucial para decidir esta pregunta es adoptar una postura en un problema al que me referiré como "la Paradoja de la Cognición" o, más simplemente, "la Paradoja".

La Paradoja es bastante fácil de identificar. De un lado, la cognición es *dura* [*hard*]: es caracterizada por las reglas de la lógica, por las reglas del lenguaje. Del otro lado, la cognición es *blanda* [*soft*]: si uno escribe las reglas, parece que la realización de esas reglas en sistemas formales automáticos (como son los programas de IA) produce sistemas que no tienen un comportamiento suficientemente fluido ni suficientemente fuerte como para constituir lo que queremos llamar inteligencia verdadera. Ésa es, de manera muy simple, la Paradoja. Al tratar de caracterizar las leyes de la cognición, somos empujados en dos direcciones diferentes: cuando focalizamos las reglas que gobiernan la competencia cognitiva de alto-nivel, somos empujados hacia representaciones y procesos simbólicos estructurados; cuando enfocamos el detalle complejo y variante del comportamiento inteligente real, somos empujados hacia descripciones numéricas estadísticas. La Paradoja podría ser llamada, de manera algo más precisa, el Dilema de la Estructura-Estadística[1] [*Structure Statistics Dilemma*]. La postura que uno adopta hacia la Paradoja influye fuertemente en el papel que pueden jugar los formalismos modeladores simbólico y conexionista. Al menos, cinco posturas dignas de atención se han adoptado sobre la Paradoja, y las presentaré rápidamente. Consideraré a cada una en su forma más pura; esas posturas extremas pueden ser vistas como caricaturas de posiciones más sutiles adoptadas realmente por los científicos cognitivos.

La primera posición que uno siempre tiene que considerar cuando se enfrenta con una paradoja es la *negación*. De hecho, ésta es probablemente la elección más común. La opción por la negación aparece de dos formas. La primera consiste en *negar lo blando*. Un nombre más respetable para esto podrá ser *racionalismo*. En esta respuesta a la Paradoja uno insiste en que la esencia de la inteligencia es lógica y sigue reglas, todo lo demás no es esencial. Esto puede ser identificado como

1. Para discusiones relacionadas, ver, por ejemplo, Gerken y Bever, 1986; Greeno, 1987.

la motivación dentro de la noción de la competencia ideal en lingüística (Chomsky, 1965), en la que la conducta blanda y la variabilidad en el comportamiento son vistos como mero ruido. El hecho de que haya una regularidad enorme en ese ruido tiene que ser ignorado, al menos en la versión más pura de esta postura.

La otra postura consiste obviamente en *negar lo duro*. De acuerdo con esta visión, el seguir reglas es verdaderamente una característica del comportamiento *novato*, no experto; la esencia de la inteligencia real es su *evasión* del seguimiento-de-reglas [*rule-following*] (Dreyfus y Dreyfus, 1986). Ciertamente, algunos de los partidarios más fuertes de esta posición son conexionistas que alegan que en la cognición "no hay reglas".

Si uno rechaza las opciones de negación puede ir al extremo opuesto, que llamaré *el cerebro dividido* [*the split brain*].[2] Según este punto de vista, la cabeza contiene una máquina blanda y una dura, y ellas se sitúan una al lado de la otra. Esta respuesta a la Paradoja está involucrada cuando se habla de sistemas que tienen "módulos conexionistas" y "módulos basados-en-reglas", y alguna clase de comunicación entre ellos. Existe el cerebro derecho conexionista haciendo el procesamiento blando [*squishy*] y el cerebro izquierdo, a la von Neumann, haciendo el procesamiento duro basado-en-reglas. En vez de "el cerebro dividido", esta escena de una casa dividida, a derecha e izquierda trabajando lado a lado a pesar de sus profundas diferencias, podría ser llamada mejor por su nombre francés: *cohabitation*.

Presumiblemente, los partidarios de esta respuesta sienten que le están dando el mismo peso a los dos lados de la Paradoja. ¿Pero atrapa esta respuesta verdaderamente toda la fuerza de la Paradoja? En el cerebro dividido, hay una *línea dura* que rodea y aísla la blandura y no hay una línea blanda que demarque la dureza. La blandura es cuidadosamente escondida en una arquitectura global caracterizada por una distinción dura entre procesamiento duro y procesamiento blando. La fuerza completa de la Paradoja insiste en que los aspectos duros y blandos de la cognición están tan íntimamente entrelazados que semejante distinción dura no es viable. Para no mencionar el serio problema de obtener que los dos tipos de sistemas cooperen íntimamente en tanto hablan lenguajes tan diferentes.

El tercer enfoque de la Paradoja es el *enfoque difuso* [*fuzzy*

2. Para resultados empíricos relevantes a este tema, en alguna medida decepcionantes, ver Bever, Carrithers y Townsend, 1987.

approach] (Gupta, Ragade y Yager, 1979). Aquí, la idea básica es tomar una máquina dura y revestir sus partes con blandura. Uno toma un sistema basado-en-reglas para hacer un diagnóstico médico y asigna un número a cada regla que dice cuán cierta es la inferencia (Shortliffe, 1976; Zadeh, 1975, 1983); o uno toma un conjunto y a cada miembro le asigna un número que dice a cuántos miembros del conjunto representa [ese número] (Zadeh, 1965). En esta respuesta a la Paradoja, la blandura es definida como grados de dureza. Uno toma la ontología del problema que proviene del enfoque duro y fija números a todos los elementos de esa ontología, en vez de reconceptualizar la ontología de una manera novedosa que refleje de modo intrínseco la blandura en el sistema.

Sobre tales bases ontológicas, el cuarto enfoque comienza a ser bastante más sofisticado. Según este punto de vista, la máquina cognitiva es en el fondo una máquina dura; fundamentalmente, todo trabaja en base a reglas, pero la máquina es tan compleja que *aparece blanda* cuando uno la observa desde un nivel superior. *La blandura emerge de la dureza*. Esta respuesta a la Paradoja está implícita en comentarios como,

> bien, quizá, mi sistema experto *es* frágil, pero eso ocurre porque es un sistema de juguete con sólo 10.000 reglas... si tuviera los recursos construiría el sistema *real* con 10^{10} reglas y sería tan inteligente como el experto humano.

Con otras palabras, si hay suficientes reglas duras dando vueltas en el sistema, la conducta fluida será una propiedad emergente.

En términos de niveles de descripción, éste es el panorama. Hay un nivel de descripción en el cual el sistema cognitivo es duro: el nivel inferior. Y hay un nivel de descripción en el cual es blando: el nivel superior. Éste es el sentido en el cual este enfoque se vuelve más sofisticado: usa *niveles de análisis* para reconciliar los lados duro y blando de la Paradoja.

Cabe preguntar si este enfoque llega a funcionar. El esfuerzo por producir sistemas construidos con un gran número de reglas duras a partir de la fragilidad que es intrínseca a tales reglas hace ya tiempo que se lleva a cabo. Que los éxitos parciales constituyan una base para el optimismo o el pesimismo es difícil de decidir.

El quinto y último enfoque que quiero considerar es uno acerca del que he argumentado (Smolensky, 1988a), que constituye la base del tratamiento apropiado del conexionismo. Según este punto de vista, que

he denominado el enfoque *subsimbólico* [*subsymbolic*], el sistema cognitivo es fundamentalmente una máquina blanda que es tan compleja que a veces parece dura, cuando es vista desde niveles superiores. Como en el enfoque anterior, la Paradoja es encarada mediante dos niveles de análisis, pero ahora es el nivel inferior el que es blando y el nivel superior el que es duro: ahora *la dureza emerge de la blandura*.

Habiendo presentado estas cinco respuestas a la Paradoja, podemos ver ahora por qué la decisión de adoptar un formalismo computacional simbólico o uno conexionista está enraizada en una postura acerca de la Paradoja. La cuestión es si [cabe] asumir un formalismo que *dé por sentadas* las características del lado duro de la Paradoja, o uno que dé por sentadas las características del lado blando. Si uno decide no combinar ambos formalismos (*cohabitación*) sino tomar uno como el fundamental, entonces, cualquiera sea el camino que uno tome, tiene que *ignorar* el otro lado, o bien *construirlo* en el formalismo que ha elegido.

Entonces, ¿cuáles son las motivaciones posibles para tomar el lado blando como el substrato fundamental sobre el cual construir el duro, cualesquiera que sean los aspectos duros de la cognición que necesiten ser construidos? He aquí algunas razones para dar al lado blando prioridad en ese sentido.

- Un enfoque fundamentalmente blando es interesante si uno ve la *percepción*, más que la *inferencia lógica*, como el soporte de la inteligencia. En la aproximación subsimbólica, la base fundamental de la cognición es vista como categorización y procesos perceptuales de esa clase.
- En un comportamiento cognitivo global, la dureza parece más la excepción que la regla. Eso corta ambos lados, por supuesto. La opción de la negación está siempre abierta para decir que lo que verdaderamente caracteriza a la inteligencia es sólo el 3 % que no es blando, y que esto es lo que debería preocuparnos.
- Un argumento evolucionista dice que el lado duro de la Paradoja cognitiva se desarrolla más tarde, por encima del lado blando y que la ontogenia teórica de uno tiene que recapitular la filogenia.
- Comparado con los enfoques simbólicos basados en reglas, es mucho más fácil ver cómo la clase de sistemas blandos, que los modelos conexionistas representan, podrían ser implementados en el sistema nervioso.
- Si uno va a basar toda la solución de la Paradoja en la emergencia de un tipo de computación sobre el otro, entonces se vuelve

crucialmente importante que seamos capaces de analizar las propiedades de nivel superior del sistema de nivel inferior. Que la matemática que gobierna las redes conexionistas pueda ser analizada por propiedades emergentes parece ser una apuesta considerablemente mejor que el que los sistemas basados en reglas extremadamente complejas sean analizables por sus propiedades emergentes. La tarea de analizar las propiedades emergentes de los sistemas conexionistas está estrechamente relacionada con los tipos tradicionales de análisis de los sistemas dinámicos en física; esto ya ha dado señales de que puede ser en última instancia exitoso.

- Por último, el lado duro ha tenido prioridad por varias décadas, con resultados decepcionantes. Es tiempo de dar al lado blando unas pocas décadas para producir resultados decepcionantes por sus propios medios.

La decisión de adoptar un enfoque fundamentalmente blando y construir un nivel duro sobre él, tiene costos serios, como lo señala con algún detalle Kirsh (1987). El poder de los símbolos y de la computación simbólica no le es dado a uno gratuitamente; uno tiene que construirlos a partir del material blando, y esto es realmente muy difícil. Hasta ahora, no sabemos cómo llevarlo a cabo. Como Kirsh señaló, si uno no tiene símbolos en el sentido usual, no está claro que pueda hacer frente a un número de problemas. La crítica de Fodor y Pylyshyn es básicamente una enunciación del mismo tipo general: que el precio que uno tiene que pagar al ser conexionista es el fracaso en dar cuenta de ciertas regularidades del lado duro, regularidades que el formalismo simbólico le da a uno prácticamente gratis.

Si la fuerza de tales críticas significa que el conexionismo no ha llegado *aún* a proveer las capacidades de la computación simbólica como para hacer justicia al lado duro de la Paradoja, creo personalmente que están en lo correcto. Adoptar la postura subsimbólica en la Paradoja equivale a contraer una deuda enorme, una deuda que escasamente ha comenzado a ser pagada.

Por otra parte, si se acepta que la fuerza de tales críticas significa que el conexionismo no podrá nunca llegar a proveer las capacidades de la computación simbólica sin implementar meramente el enfoque simbólico, entonces, como argumentaré en el resto de este artículo, creo que tales críticas deben ser rechazadas.

¿Cuáles son los beneficios de seguir el enfoque subsimbólico de la

Paradoja? ¿Por qué vale la pena contraer esa deuda enorme? En mi opinión, la justificación principal es que si tenemos éxito en construir símbolos y manipulación simbólica a partir del "conectoplasma" [*connectoplasm*] entonces tendremos una explicación de *dónde provienen los símbolos y la manipulación simbólica*, y eso vale el riesgo y el esfuerzo; y lo vale en grado sumo. Con suerte, tendremos aun una explicación de cómo el *cerebro* construye la computación simbólica. Pero incluso, si no obtenemos eso directamente, será la primera teoría acerca de cómo obtener símbolos a partir de alguna cosa que se asemeje remotamente al cerebro, y eso ciertamente será útil (en verdad, argumentaría, crucial) para imaginarnos cómo el cerebro realmente lo hace.

Otro pago potencial es contar con una manera de explicar *por qué* aquellos aspectos de la cognición que exhiben dureza deberían exhibir dureza: por qué el área de la dureza cae donde cae; por qué está limitada como está; por qué el enfoque simbólico triunfa donde triunfa y fracasa donde fracasa.

Por último, por supuesto, si el enfoque subsimbólico triunfa, tendremos en verdad una solución unificada de la Paradoja: no habrá negación de la mitad del problema ni una división profunda del cerebro.

Ya se pueden visualizar contribuciones que conducen a estos últimos resultados. El enfoque conexionista está produciendo conceptos y técnicas nuevos para abarcar las regularidades en el comportamiento cognitivo, tanto en el nivel inferior en el que el marco teórico conexionista se aplica naturalmente como en el nivel superior en el que los enfoques simbólicos son importantes. (Para estudios recientes, ver McClelland, Rumelhart y el Grupo de Investigación PDP, 1986; Rumelhart, McClelland y el Grupo de Investigación PDP, 1986; Smolensky, en prensa). El repertorio teórico de la ciencia cognitiva y computacional se está enriqueciendo con las nuevas concepciones acerca de cómo puede ser hecha la computación.

En lo que respecta a dónde estamos en cuanto a la consecución de los objetivos últimos, en mi opinión, lo que tenemos son técnicas interesantes y sugerencias promisorias. Nuestra posición habitual en la historia intelectual de la computación conexionista, según mi punto de vista, puede ser expresada por esta analogía:

$$\frac{\text{comprensión ordinaria de la computación conexionista}}{\text{comprensión ordinaria de la computación simbólica}} :: \frac{\text{Aristóteles}}{\text{Turing}}$$

Estamos, de alguna manera, aproximándonos a la posición de Aristóteles en el desarrollo intelectual de este nuevo enfoque computacional. Si hay conexionistas entusiastas que piensan que realmente podemos modelar la cognición desde tal posición, están, me temo, muy equivocados. Y si no podemos ir de Aristóteles a (al menos) Turing en nuestra comprensión de la computación subsimbólica, no nos vamos a acercar a la cognición real mucho más de lo que estamos ahora.

Un comentario final antes de abordar la crítica de Fodor y Pylyshyn. La explicación dada aquí en relación con la elección de un marco teórico conexionista para la paradoja duro/blando, arroja alguna luz sobre la pregunta hecha a menudo por los observadores de la sociología del conexionismo: "¿Por qué el club de entusiastas del conexionismo incluye un surtido tan extraño de personas?". Al menos, en la lectura amable de esta pregunta, "surtido extraño" se refiere a un grupo filosóficamente heterogéneo de científicos cognitivos cuyos puntos de vista tienen en común poco más que un rechazo del paradigma simbólico predominante. Mi respuesta a esta pregunta es que la prioridad de lo duro ha hecho a muchas personas muy infelices durante mucho tiempo. El fracaso en hacer justicia al lado blando de la Paradoja por parte de los enfoques formales predominantes de los procesos cognitivos ha hecho que personas de muy distintas perspectivas se sientan alienadas por el esfuerzo. Asignándole a lo blando la posición prioritaria, haciendo de ello la base del formalismo, el conexionismo ha dado a muchas personas un apoyo formal que no tuvieron. Y por eso *deberían* estar felices.

En este punto, "conexionismo" refiere más a un formalismo que a una teoría. Así, no es apropiado parafrasear la pregunta del parágrafo anterior como "¿Qué tipo de teoría tendría como adherentes a tan surtido grupo de personas?". No es realmente una pregunta acerca de una teoría, es en realidad una pregunta acerca de qué tipo de *formalismo* permite a las personas con diferentes teorías, decir lo que necesitan decir.

Habiendo expuesto mi posición de que la comprensión de la elección de un formalismo conexionista involucra considerar posturas alternativas con respecto a la Paradoja de la Cognición, procedo ahora a considerar la crítica de Fodor y Pylyshyn bajo esa luz.

2. *Fodor y Pylyshyn acerca de la estructura constitutiva de los estados mentales*

He aquí una rápida síntesis del argumento central de Fodor y Pylyshyn (1988).

(1) Los pensamientos tienen una estructura compuesta.

Con esto quieren decir cosas como: el pensamiento de que *Juan ama a la muchacha* no es atómico; es un estado mental compuesto, construido a partir de pensamientos acerca de *Juan*, *ama*, *a*, y *la muchacha*.

(2) Los procesos mentales son sensibles a esa estructura compuesta.

Por ejemplo, a partir de cualquier pensamiento de la forma *p y q* —sin importar lo que *p* y *q* sean— podemos deducir *p*.

Fodor y Pylyshyn elevan (1) y (2) al *status* de una definición de la Perspectiva Clásica de la Cognición y dicen que esto es lo que los conexionistas han cuestionado. Más adelante argumentaré que están equivocados, pero continuemos ahora con su argumento.

Habiendo identificado las afirmaciones (1) y (2) como definitorias de la Perspectiva Clásica, Fodor y Pylyshyn sostienen que hay argumentos decisivos a su favor. [Ellos admiten que esos argumentos son una reedición moderna de los '80, una versión en colores de una película que fue exhibida en blanco y negro algún tiempo atrás, en la que la palabra "behaviorismo" fue reemplazada por "conexionismo"]. Los estados mentales tienen, de acuerdo con esos argumentos, las propiedades de productividad, sistematicidad, composicionalidad y coherencia inferencial. Sin entrar en todos esos argumentos, permítaseme simplemente decir que para los propósitos presentes estoy dispuesto a aceptar que son suficientemente convincentes como para justificar la conclusión de que (1) y (2) tienen que ser tomadas muy seriamente. Cualesquiera que sean las inclinaciones de otros conexionistas, esos argumentos y otros argumentos relacionados me convencen que negar lo duro es un error. No me convencen de que deba negar lo blando, tampoco, presumiblemente, lo intentan.

Veamos ahora el análisis del conexionismo que ofrecen Fodor y Pylyshyn. Ellos aseveran que en el conexionismo (estándar), *todas las representaciones son atómicas*; los estados mentales no tienen estructura compuesta, violando (1). Además, aseveran que *el procesamiento conexionista* (estándar) *es la asociación*, y es sensible sólo a la *estadística*, no a la *estructura*, violando (2). En consecuencia concluyen que el conexionismo (estándar) es de manera maximal no Clásico: viola ambos prin-

cipios definitorios. Por ende, el conexionismo es derrotado por los argumentos decisivos en favor de la Perspectiva Clásica. ¿Qué es lo que hace que Fodor y Pylyshyn digan que las representaciones conexionistas son atómicas? La segunda figura de su artículo (pág. 16) lo dice todo (aparece aquí como la figura 1). Se supone que esa red ilustra el enfoque conexionista estándar de la inferencia de *A* y *B* a *A* y a *B*. Es cierto que Ballard y Hayes escribieron un artículo (Ballard y Hayes, 1984) acerca del uso de redes conexionistas para la resolución de teoremas, en el que aparecían redes como ésta. Sin embargo, es un serio error ver esto como el enfoque conexionista paradigmático para inferencias humanas de ese tipo. Esta clase de representación conexionista *ultra-local*, en la que proposiciones enteras están representadas por nodos individuales, está lejos de ser lo típico en los modelos conexionistas, y ciertamente, no tiene que ser tomada como *definitoria* del enfoque conexionista.

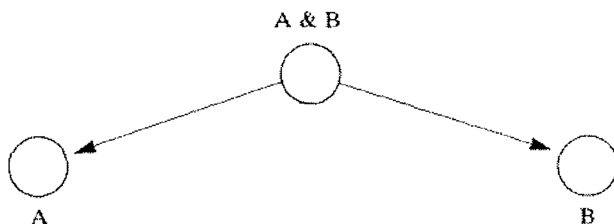


Figura 1: La red de Fodor y Pylyshyn

Mi contraargumento central contra Fodor y Pylyshyn comienza con la afirmación de que cualquier crítica al enfoque conexionista tiene que considerar las consecuencias de usar *representaciones distribuidas*, en las que las representaciones de entidades conceptuales de alto nivel, tales como las proposiciones, están distribuidas en muchos nodos y los mismos nodos participan simultáneamente en la representación de muchas entidades. Su respuesta, en la Sección 2.1.3. (pág. 19), es la siguiente. La cuestión representación local/representación distribuida tiene que ver (según ellos asumen) con que si cada uno de los nodos de la figura 1 se refiere a algo complicado y de nivel inferior (el caso distribuido), o no (el caso local). Pero, sostienen, esa cuestión es irrelevante porque atañe a una cuestión de *nivel* '*entre*' [*a between level issue*] y la composicionalidad de los estados mentales es una cuestión de *nivel* '*dentro de*' [*a within level issue*].

Mi respuesta es que están en lo correcto con respecto a que la composicionalidad es una cuestión de nivel 'dentro de' y que la distinción distribuido/local es una cuestión de nivel 'entre'. Su argumento supone que debido a esa diferencia una cuestión no puede influenciar a la otra. Pero esto es una falacia. Se asume que en las representaciones distribuidas la relación de nivel 'entre' no puede tener ninguna consecuencia en la estructura de *nivel 'dentro de'* de las relaciones entre las representaciones de *A* y *B* y la representación de *A*. Y esto es simplemente falso. Hay implicaciones de las representaciones distribuidas para la composicionalidad, las que introduciré en el resto de esta sección por medio de un ejemplo extenso. En particular, resultará que la figura 1 no es más relevante para un enfoque conexionista distribuido de la inferencia que para un enfoque simbólico. En el caso hiper-local, la figura 1 es relevante y su crítica se mantiene; en el caso distribuido, la figura 1 es una caracterización falsa del enfoque conexionista y su crítica no da en el blanco. Además, resultará que un análisis válido del caso distribuido real, basado en las sugerencias del propio Pylyshyn, conduce a la conclusión completamente opuesta: los modelos conexionistas que usan representaciones distribuidas describen los estados mentales con un tipo relevante (nivel 'dentro de') de estructura constitutiva.

Antes de desarrollar este contraargumento, permítaseme resumir el punto básico del artículo de Fodor y Pylyshyn. Dado que ellos creen que el conexionismo estándar es fatalmente defectuoso, propugnan que los conexionistas persigan en cambio un conexionismo no estándar. Los conexionistas deberían admitir los principios (1) y (2); deberían aceptar la perspectiva clásica y deberían diseñar sus redes para que sean implementaciones de las arquitecturas clásicas. Aquí, la lógica implícita es que los modelos conexionistas que respeten (1) y (2) tienen que ser necesariamente implementaciones de una arquitectura clásica; ésta es su segunda gran falacia, a la que volveré en la Sección 3. Fodor y Pylyshyn sostienen que el conexionismo debería ser usado para implementar arquitecturas clásicas, y que una vez que haya hecho eso, el conexionismo no proporcionará una arquitectura cognitiva nueva sino una implementación para la arquitectura cognitiva vieja; que lo que el conexionismo puede proporcionar, por ende, no es un nuevo paradigma para la ciencia cognitiva, sino más bien alguna información nueva sobre "la ciencia de la implementación" o, posiblemente, la neurociencia.

Si los conexionistas siguieran la estrategia de implementación que Fodor y Pylyshyn propugnan, creo que estas consecuencias concernientes a la arquitectura clásica ciertamente se seguirían. Pero no creo que

de aceptar (1) y (2) se siga que las redes conexionistas tienen que ser implementaciones. En la Sección 3 argumento que los conexionistas pueden aceptar de manera consistente (1) y (2), y rechazar el enfoque implementacionista que Fodor y Pylyshyn propugnan.

Por ahora, el objetivo es mostrar que los modelos conexionistas que usan representaciones *distribuidas* adscriben a los estados mentales la clase de estructura composicional que (1) demanda, en oposición a la conclusión de Fodor y Pylyshyn que se basa en que la red de la figura 1 encarna una representación hiper-local.

Unidades	*Microcaracterísticas*
•	recipiente vertical
•	líquido caliente
○	vaso contactando madera
•	superficie curva de porcelana
•	olor a quemado
•	líquido marrón contactando porcelana
•	superficie curva de porcelana
○	objeto de plata oblongo
•	asa del tamaño del dedo
•	líquido marrón con lados curvos y fondo

Figura 2. Representación de *taza* con *café*

Mi argumento consiste principalmente en llevar a cabo un análisis sugerido por el mismo Zenon Pylyshyn en el Encuentro de Ciencia Cognitiva de 1984 en Boulder. Se produjo una suerte de debate sobre el conexionismo entre Geoffrey Hinton y David Rumelhart, de un lado, y Zenon Pylyshyn y Kurt Van Lehn, del otro lado. Al discutir la naturaleza de las representaciones conexionistas, Pylyshyn le pregunta a Rumelhart: "Mire, ¿pueden representar una taza de café en esas redes?". La respuesta de Rumelhart fue "Seguro". Entonces, Pylyshyn continuó: "¿Y pueden representar en ella una taza sin café?". Advirtiendo que la trampa se acercaba, Rumelhart dijo "Sí", punto en el que Pylyshyn exclamó : "Ah, bien, la diferencia entre las dos es sólo la representación de *café* y usted ha construido una representación de *taza con café,* combinando una representación de *taza* con una representación de *café*".

Desarrollemos con exactitud la construcción sugerida por Pylyshyn y veamos a qué conclusión nos lleva. Tomaremos una representación

distribuida de *taza con café* y abstraeremos de ella una representación distribuida de *taza sin café* y llamaremos a lo que quedó "la representación conexionista de *café*".

Para generar esas representaciones distribuidas usaré un conjunto de "microcaracterísticas" [*microfeatures*] (Hinton, McClelland y Rumelhart, 1986) que no son muy micro, pero esto es lo que siempre pasa cuando uno trata de crear ejemplos que puedan ser intuitivamente entendidos en una exposición no técnica. Esas microcaracterísticas se muestran en la figura 2.

La figura 2 muestra una representación distribuida de *taza con café*: un patrón de actividad en el que las unidades que están activas (en negro) son las que corresponden a las microcaracterísticas presentes en la descripción de una taza que contiene café. Obviamente, ésta es una representación cruda, cercana al nivel sensorial, pero de nuevo, esto ayuda a que el ejemplo sea más intuitivo: no es esencial.

Dada la representación de *taza con café* expuesta en la figura 2, Pylyshyn nos sugiere abstraer la representación de *taza sin café*. La representación de *taza sin café* se muestra en la figura 3, y la figura 4 muestra el resultado de abstraerla de la representación de *taza con café*.

¿Qué produce este procedimiento como "la representación conexionista de *café*"? A partir de la figura 4 tenemos olor a quemado y líquido marrón caliente con superficies de lados curvos y fondo que contacta porcelana. Ésta es ciertamente una representación de *café*, pero en un contexto muy particular: el contexto provisto por *taza*.

¿Qué significa esto en cuanto a la conclusión de Pylyshyn de que "la representación conexionista de *taza con café* es sólo la representación de *taza sin café* combinada con la representación de *café*"? ¿Qué está involucrado en la combinación conjunta de las representaciones de las figuras 3 y 4 para formar la de la figura 2? Ensamblamos la representación de *taza con café* a partir de la representación de una *taza*, y una representación de *café*, pero es una combinación bastante extraña. Se ha obtenido también una representación de la interacción de la taza con café, como *líquido marrón contactando porcelana*. De este modo, la representación compuesta está construida a partir de *café* extraída de la situación *taza con café*, junto con taza extraída de la situación *taza con café*, junto con su interacción.

Así, la estructura composicional está allí, pero está allí en un sentido *aproximativo*. *No* es equivalente a tomar una representación de *café* independiente del contexto y una representación de *taza* independiente del contexto —y ciertamente no es equivalente a tomar una represen-

Unidades	*Microcaracterísticas*
•	recipiente vertical
○	líquido caliente
○	vaso contactando madera
•	superficie curva de porcelana
○	olor a quemado
○	líquido marrón contactando porcelana
•	superficie curva de porcelana
○	objeto oblongo de plata
•	asa del tamaño del dedo
○	líquido marrón con lados curvos y fondo

Figura 3. Representación de *taza sin café*

Unidades	*Microcaracterísticas*
○	recipiente vertical
•	líquido caliente
○	vaso contactando madera
○	superficie curva de porcelana
•	olor a quemado
•	líquido marrón contactando porcelana
○	superficie curva de porcelana
○	objeto oblongo de plata
○	asa del tamaño del dedo
•	líquido marrón con lados curvos y fondo

Figura 4. Representación de *café*

tación contextualmente independiente de la relación *en* o *con*— uniéndolas en una estructura simbólica, concatenándolas para formar las clases de estructuras composicionales sintácticas que Fodor y Pylyshyn piensan que las redes conexionistas deberían implementar.

Para extender este punto aún más, consideremos la representación de *café* una vez que la *taza* ha sido abstraída. Ésta, sugiere Pylyshyn, es la representación conexionista de *café*. Pero como ya hemos observado, ésta es realmente una representación de café en el contexto particular de estar dentro de una taza. De acuerdo con la fórmula de

Pylyshyn, para obtener la representación conexionista de *café* debería ser posible, en principio, tomar la representación conexionista de *lata con café* y abstraer a partir de ella la representación conexionista de *lata sin café*. ¿Qué ocurriría si realmente lo hiciéramos? Obtendríamos una representación de gránulos marrones semejantes a tierra con aroma a quemado apiladas en una forma cilíndrica, junto con gránulos contactando hojalata. Ésta es la representación conexionista de *café* que obtenemos cuando empezamos con *lata con café* en vez de *taza con café*. O podríamos empezar con la representación de *árbol con café* y abstraer *árbol sin café*. Obtendríamos una representación conexionista de *café*, que sería una representación de granos marrones suspendidos en el aire en forma graciosa. O de nuevo, podríamos empezar con *hombre con café* y obtener todavía otra representación conexionista de *café*: una representación casi similar a la representación entera de *taza con café* a partir de la cual extrajimos nuestra primera representación de *café*.

El punto es que la representación de *café* que extrajimos de la construcción que comenzó con *taza con café*, conduce a una representación diferente de *café* de la que obtuvimos de otras construcciones que tienen un *status* a priori equivalente. Esto quiere decir que si uno quiere hablar acerca de la representación conexionista de *café* en este esquema distribuido, uno tiene que hablar acerca de una *familia de patrones de actividad distribuida* [*family of distributed activity patterns*]. Lo que une a todas estas representaciones particulares de *café* no es otra cosa que un *parecido de familia*.

La primera moraleja que quiero extraer de esta historia del *café*, es ésta: a diferencia del caso hiper-local de la figura 1, con representaciones distribuidas, las representaciones complejas [*complex representations*] *están* compuestas por las representaciones de los componentes. La relación de constitutividad es aquí una relación de *nivel* '*dentro de*', como Fodor y Pylyshyn demandan: el patrón o *vector* que representa *taza con café* está compuesto por un *vector* que puede ser identificado como una representación distribuida de *taza sin café* junto con un *vector* que puede ser identificado como una representación distribuida particular de *café*. Al caracterizar los vectores constituyentes del vector que representa el compuesto, *no* nos concierne el hecho de que el vector que representa *taza con café* sea un vector comprendido por la actividad de unidades de microcaracterísticas individuales. La relación de *nivel* '*entre*' del vector y sus elementos numéricos individuales, *no* es la relación de constitutividad, y entonces la sección 2.1.4 (págs. 19-28) de

Fodor y Pylyshyn (1988) es irrelevante; ellos apuntan allí a un error que no ha sido cometido.

La segunda moraleja es que la relación de constitutividad entre las representaciones distribuidas es importante para el análisis de los modelos conexionistas y para explicar su conducta, pero *no* es una parte del mecanismo causal dentro del modelo. Para procesar el vector que representa *taza con café*, la red no tiene que descomponerlo en componentes. Para procesar, lo que importa es la relación de *nivel 'entre'*, no la relación de nivel 'dentro de'. El procesamiento del vector que representa *taza con café* está determinado por las actividades numéricas individuales que componen el vector: es sobre esas actividades de nivel inferior que los procesos son definidos. Así, el hecho de que haya una considerable arbitrariedad en la manera en que los constituyentes de *taza con café* son definidos, no introduce ambigüedades en la manera en que la red procesa esa representación: las ambigüedades sólo existen para nosotros que analizamos el modelo y tratamos de explicar su comportamiento. Cualquier definición particular de constitutividad que nos dé un poder explicativo, es una definición válida de la constitutividad; la falta de unicidad no es un problema.

Esto nos conduce directamente a la tercera moraleja: que la descomposición de estados compuestos en sus componentes no es precisa ni está definida unívocamente. La noción de constitutividad es importante pero probablemente los intentos de formalizarla involucrarán, de modo crucial, una *aproximación*. Tal como se discutió con alguna extensión en Smolensky (1988a), éste es el caso típico: las nociones que provienen de la computación simbólica proporcionan instrumentos importantes para construir enfoques de nivel superior del comportamiento de modelos conexionistas que usan la representación distribuida; pero estas nociones proporcionan enfoques aproximados, imprecisos.

Esto nos conduce a la cuarta moraleja: que mientras que las redes conexionistas que usan representaciones distribuidas describen estados mentales con el tipo de constitutividad requerida por (1), *no* proporcionan una implementación literal de un lenguaje sintáctico del pensamiento. La dependencia contextual de los componentes, las interacciones que tienen que ser acomodadas cuando están combinados, la inhabilidad para identificarlos de manera única, precisa, la necesidad de tomar seriamente la noción de que la representación de *café* es una colección de vectores enlazados por una familia de parecidos, todo esto implica que la relación entre la constitutividad conexionista y la constitutividad sin-

táctica no es de implementación literal. En particular, sería absurdo sostener que aun si el relato conexionista fuera correcto, entonces esto no tendría implicaciones para la arquitectura cognitiva, que meramente llenaría detalles en el nivel inferior sin implicaciones importantes para la perspectiva del nivel superior.

Todas estas conclusiones toman en cuenta a (1) sin tomar en cuenta, explícitamente, a (2). Tomar en cuenta a (2) de manera apropiada está más allá del alcance de este artículo. Hasta cierto punto, está más allá del alcance del conexionismo corriente. Permítaseme simplemente señalar que el Dilema Estructura/Estadística tiene una solución posible atractiva que el enfoque conexionista está perfectamente en condiciones de abordar: *la mente es una máquina sensible a la estadística que opera sobre representaciones numéricas sensibles a la estructura*. Los argumentos previos han mostrado que las representaciones distribuidas poseen relaciones de constitutividad, y que analizadas adecuadamente, esas representaciones pueden ser vistas como codificadoras de estructura. Extender esto para lidiar con toda la complejidad de los tipos de estructuras ricas implicadas en los procesos cognitivos complejos, es un problema de investigación que ha sido abordado con algún éxito pero que todavía no ha sido concluido de modo definitivo (ver Smolensky, 1987, y Sección 3, abajo). Una vez que tenemos la información compleja estructurada, representada en patrones numéricos distribuidos, los procesos sensibles a la estadística pueden proceder a analizar las regularidades estadísticas de una manera completamente sensible a la estructura. Que tales procesos puedan hacer frente a toda la fuerza del Dilema Estructura-Estadística, es un asunto que quedará por algún tiempo como una pregunta abierta.

La conclusión entonces, es que los modelos distribuidos *pueden* satisfacer (1) y (2). Que (1) y (2) puedan ser satisfechos al punto de proveer un enfoque adecuado para cubrir la *totalidad de las demandas de la modelización cognitiva* es por supuesto una cuestión empírica abierta: tal como es para el enfoque simbólico satisfacer (1) y (2). Del mismo modo, los modelos distribuidos conexionistas *no* equivalen a una implementación de las instanciaciones simbólicas de (1) y (2) que Fodor y Pylyshyn propugnan.

Antes de recapitular, me gustaría volver a la figura 1. ¿En qué sentido se puede decir que la figura 1 describe la relación entre las representaciones distribuidas de *A & B* y las representaciones distribuidas de *A* y *B*? El ejemplo del *café* intentó mostrar que las representaciones distribuidas son, en un sentido aproximado pero relevante para la explica-

ción, parte de la representación del compuesto. De este modo, en el caso distribuido, la relación entre el nodo de la figura 1 rotulado *A & B* y los otros nodos, es una especie de relación todo/parte. Un mecanismo de inferencia que tome como *input* al vector que representa *A & B* y produzca como *output* el vector que representa *A*, es un mecanismo que extrae una parte de un todo. Y en ese sentido, no es diferente de un mecanismo de inferencia simbólico que toma la estructura sintáctica *A & B* y extrae de ella el componente sintáctico *A*. Los mecanismos conexionistas para hacer esto son, por supuesto, completamente diferentes de los mecanismos simbólicos, y la naturaleza aproximada de la relación todo/parte le da a la computación conexionista características completamente diferentes: no tenemos, simplemente, una implementación nueva de la antigua computación.

Está claro que así como la figura ofrece un esquema crudo del proceso simbólico de pasar de *A & B* a *A*, un esquema que usa los rótulos para codificar estructuras internas ocultas dentro de los nodos, *exactamente lo mismo es verdad para el caso del conexionismo distribuido*. En el caso distribuido, como en el caso simbólico, los enlaces de la figura 1 son esquemas crudos de procesos complejos y no canales causales simples que pasan actividad desde el nodo más alto a los nodos más bajos. Semejante relato causal se aplica sólo al caso conexionista hiper-local, que sirve aquí como el proverbial "hombre imaginario" [*straw man*].

Permítaseme ser claro: hasta donde sé, no hay un modelo distribuido conexionista de la clase de inferencia formal que Fodor y Pylyshyn tienen en mente aquí. Tal inferencia formal está localizada en el extremo remoto del lado duro de la Paradoja y no es, en ese punto, un proceso cognitivo (o abstracción de él) que el formalismo conexionista pueda decir que lo ha construido sobre su sustrato blando. Pero en el fondo, la crítica de Fodor y Pylyshyn gira en torno a la estructura constitutiva de los estados mentales; la inferencia formal es sólo uno de los escenarios en los cuales se ve la importancia de esa estructura constitutiva. Así, la discusión precedente acerca de la estructura constitutiva de las representaciones distribuidas conduce al corazón de sus críticas, aun cuando no dispongamos de un enfoque conexionista bien desarrollado de la inferencia formal.

Permítaseme resumir en este punto el panorama global. Obtuvimos los principios (1) y (2) y obtuvimos una instanciación simbólica de ellos en un lenguaje del pensamiento usando la constitutividad sintáctica [*syntactic constituency*]. Lo que deberían hacer los conexionistas, de acuerdo con Fodor y Pylyshyn, es tomar ese lenguaje simbólico del pen-

samiento [*symbolic language of thought*] como una descripción de nivel superior y entonces producir una implementación conexionista, en un sentido literal. Entonces, las operaciones sintácticas del lenguaje simbólico del pensamiento proporcionarían una descripción formal exacta de nivel superior.

En oposición, argumento que la perspectiva de la composicionalidad conexionista distribuida nos permite instanciar los mismos principios básicos de (1) y (2) *sin* pasar por un lenguaje simbólico del pensamiento. Yendo directamente a los modelos distribuidos conexionistas, obtenemos *instanciaciones nuevas de los principios de composicionalidad.*

Suelo creer que las descripciones simbólicas proveen descripciones aproximadas útiles de nivel superior de cómo esos modelos conexionistas computan, pero en ningún sentido esos modelos distribuidos conexionistas proveen una implementación literal de un lenguaje del pensamiento simbólico. Las aproximaciones requieren una predisposición para aceptar símbolos sensibles al contexto y componentes interaccionales presentes en estructuras composicionales, y la curiosa cuestión que surgió en el ejemplo del *café*. Si uno está dispuesto a vivir con todos esos grados de aproximación, entonces uno puede útilmente ver esas descripciones de nivel simbólico como descripciones aproximadas de alto nivel del procesamiento en una red conexionista.

La conclusión global es, entonces, que *los enfoques clásico y conexionista no difieren en si aceptan los principios* (1) *y* (2), *sino en cómo los instancian formalmente*. Para confrontar la verdadera disputa clásico/conexionista, uno tiene que estar dispuesto a descender al nivel de las instanciaciones formales particulares que ellos dieron a esos principios no formales. No descender a ese nivel de detalle es perder de vista el problema. En el enfoque clásico, los principios (1) y (2) son formalizados usando estructuras sintácticas para pensamientos y manipulación simbólica para los procesos mentales. En la perspectiva conexionista, (1) y (2) son formalizados usando representaciones vectoriales distribuidas para los estados mentales y la correspondiente noción de composicionalidad, junto con procesos mentales basados en la asociación que derivan su sensibilidad a la estructura de la sensibilidad a la estructura de las representaciones vectoriales comprometidas en esos procesos.

En términos de metodología de investigación, esto significa que la agenda para el conexionismo no debería ser el desarrollo de una implementación conexionista del lenguaje del pensamiento simbólico sino,

más bien, el desarrollo del análisis formal de las representaciones vectoriales de estructuras complejas y de las operaciones en aquellas estructuras que son lo suficientemente sensitivas a la estructura como para realizar el trabajo requerido.

En resumen: las representaciones distribuidas proveen una descripción de los estados mentales con componentes interpretables semánticamente, pero no hay un enfoque formal preciso de la construcción de compuestos a partir de componentes interpretables semánticamente, independientes del contexto. Según este enfoque, *hay* un lenguaje del pensamiento, pero sólo aproximativamente; el lenguaje del pensamiento no proporciona una base para una descripción formal exacta de la estructura o de los procesos mentales: no puede proveer un enfoque formal preciso de la arquitectura cognitiva.[3]

3. Conexionismo e implementación

En la Sección 2 argumenté que la investigación conexionista debería estar dirigida a las representaciones y procesos sensibles a la estructura, pero no a la implementación de un lenguaje del pensamiento simbólico. En esta sección quiero considerar el terreno que media entre la implementación de la computación simbólica y la ignorancia de la estructura. Muchos críticos del conexionismo parecen no entender que este terreno existe. (Para una mayor discusión sobre este punto y un mapa que localiza explícitamente ese terreno, ver Smolensky, 1988b.)

Una conclusión bastante específica de la Sección 2 fue que los cone-

3. Una pregunta pendiente [*open question*] importante es si la clase de relato que he dado sobre la *taza de café*, usando esas microcaracterísticas, se trasladará a la clase de representaciones distribuidas que las redes conexionistas reales crean para ellas mismas en sus unidades ocultas; si uno hace el análisis adecuadamente sofisticado. La resolución de este punto depende de la naturaleza (hasta ahora inescrutable) de esas representaciones para los problemas reales. La naturaleza del problema es importante, pues es perfectamente probable que las redes conexionistas desarrollen representaciones composicionales en sus unidades ocultas sólo cuando ello es ventajoso para el problema que están tratando de resolver. Como Fodor y Pylyshyn y todo el paradigma Clásico argumentan, tales representaciones composicionales son de hecho enormemente útiles para un amplio espectro de problemas cognitivos. Pero antes de que tales problemas, que tienden a ser considerablemente más sofisticados que aquellos usualmente dados a las redes conexionistas, hayan sido explorados con algún detalle por los modelos conexionistas, no sabremos realmente si las unidades ocultas desarrollarán representaciones composicionales (en el sentido aproximado discutido en este artículo) cuando "deberían" hacerlo.

xionistas necesitan desarrollar el análisis de las representaciones (vectoriales) distribuidas de estructuras compuestas y de las clases de procesos que operan sobre ellas con una necesaria sensibilidad a la estructura. De modo más general, mi caracterización de la meta de la modelización conexionista consiste en desarrollar modelos formales de los procesos cognitivos que estén basados en las matemáticas de los sistemas dinámicos que se desarrollan continuamente en el tiempo: sistemas complejos de variables numéricas gobernadas por ecuaciones diferenciales. Estas descripciones formales moran en la categoría de las matemáticas continuas más que en dependencias de las matemáticas discretas que subyacen al formalismo simbólico tradicional. Esta caracterización de la meta del conexionismo está lejos de ser universal: es completamente inconsistente con la caracterización definitoria de Feldman y Ballard (1982), por ejemplo. En Smolensky (1988a) argumento con cierta extensión que mi caracterización, denominada TAC, constituye un Tratamiento Adecuado del Conexionismo [*Proper Treatment of Connectionism*].

Un componente central del TAC es la relación hipotetizada entre los modelos conexionistas basados en una matemática continua y los modelos clásicos basados en una computación simbólica discreta. Esa relación, que aparece al pasar en el argumento de Fodor y Pylyshyn de la Sección 2, podría ser llamada el *principio de la correspondencia cognitiva* [*cognitive correspondence principle*]: cuando los sistemas conexionistas son analizados en niveles superiores, los elementos de la computación simbólica aparecen como propiedades emergentes.

La figura 5 ilustra el principio de correspondencia cognitiva. En la parte superior tenemos nociones no formales: la hipótesis central de que los principios de la cognición consisten en principios de memoria, inferencia, composicionalidad y estructura constitutiva, etc. En el argumento de Fodor y Pylyshyn, los principios no formales relevantes eran sus principios de la composicionalidad (1) y (2).

Los principios no formales del extremo superior de la figura 5 tienen ciertas formalizaciones en la categoría discreta, que se muestran en un nivel más abajo en la rama derecha. Por ejemplo, la memoria es formalizada como memoria estándar dirigida localmente [*standard location-addressed memory*] o alguna noción relacionada apropiada más sofisticada. La inferencia es formalizada en la categoría discreta como inferencia lógica, una forma particular de manipulación simbólica. Y así en más.

La agenda TAC consiste en tomar esas clases de principios cogni-

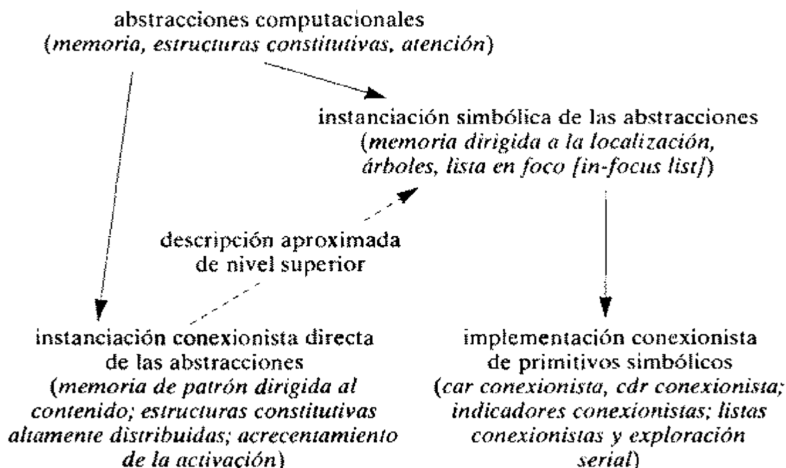


Figura 5. TAC *vs.* implementacionismo [*implementationalism*]

tivos y encontrar nuevas maneras de instanciarlos en principios formales basados en las matemáticas de los sistemas dinámicos; en la figura 5 se muestran en el nivel más bajo de la rama izquierda. El concepto de recuperación de memoria [*memory retrieval*] es reformalizado en términos de la evolución continua de un sistema dinámico hacia un punto de atracción [*point attractor*], cuya posición en el estado espacio es la memoria; uno obtiene, naturalmente, memoria dirigida al contenido [*content addressed memory*] en vez de memoria dirigida localmente (El almacenamiento de la memoria se torna una modificación del sistema para que sus atractores [*attractors*] sean localizados donde se supone que están los recuerdos; así, los principios del almacenamiento de la memoria son aun más disímiles con respecto a sus contrapartes simbólicas que los de la recuperación de memoria). Cuando reformalizamos los principios de la inferencia, el formalismo continuo conduce naturalmente a los principios de la inferencia estadística más que a la inferencia lógica. Y así en más.

El principio de correspondencia cognitiva establece que la relación general entre los principios formales conexionistas y los principios for-

males simbólicos —dado que ambos son instanciaciones de nociones comunes no formales— consiste en que si uno toma un nivel superior de análisis de lo que está sucediendo en los sistemas conexionistas, encuentra que encaja, en algún grado de aproximación, con lo que está sucediendo en el formalismo simbólico. Esta relación está indicada en la figura 5 por la flecha punteada.

Esto merece ser contrastado con el enfoque implementacionista del conexionismo que Fodor y Pylyshyn propugnan. Tal como se refleja en la figura 5 la metodología de implementación tiene que proceder desde arriba hacia abajo no directamente, *via* la rama izquierda, sino indirectamente, *via* la rama derecha; los conexionistas deberían tomar las instanciaciones simbólicas de los principios no formales y deberían encontrar maneras de implementarlas en redes conexionistas.

La metodología TAC es contrastada no sólo con el enfoque implementacionista, sino también con el eliminativista. En los términos de estas consideraciones metodológicas, el eliminativismo tiene una forma fuerte y una débil. La forma débil aboga tomar de la rama izquierda de la figura 5, ignorando completamente las formalizaciones simbólicas, en la creencia de que las nociones simbólicas confundirán más bien que iluminarán nuestros intentos por comprender la computación conexionista. La posición eliminativista fuerte sostiene que es un error tomar aun los principios no formales, en el extremo superior de la figura 5 como un punto de partida para pensar acerca de la cognición; por ejemplo, que es mejor abordar una estrategia ciega abajo-arriba en la que los principios conexionistas de bajo nivel son tomados de la neurociencia y ver luego adónde nos conducen, sin ser influidos prejuiciosamente por nociones arcaicas precientíficas tales como las del extremo superior de la figura 5.

Al rechazar las posiciones implementacionista y eliminativista, el TAC ve las descripciones conexionistas como reduciendo y explicando los enfoques simbólicos. Las descripciones conexionistas sirven para refinar los enfoques simbólicos, para reducir el grado de aproximación requerida, para enriquecer las nociones computacionales a partir del mundo simbólico y discreto, para completarlas con nociones de la computación continua. Principalmente, esto se realiza descendiendo a un nivel inferior de análisis, prestando atención a la microestructura implícita en esas clases de operaciones simbólicas.

Llamo a esto el principio de la correspondencia cognitiva porque creo que en el desarrollo de la microteoría de la cognición tiene un papel análogo al papel que el principio de correspondencia cuántico

jugó en el desarrollo de la microteoría en la física. El argumento a partir de la física encarna la estructura de la figura 5 de manera bastante directa. Hay ciertos principios físicos que forman un arco sobre los formalismos clásico y cuántico: las nociones de espacio y tiempo y los principios de invariancia asociativa, los principios de energía y conservación del *momentum*, las leyes de fuerza, etcétera. Los principios del extremo superior de la figura 5 son instanciados de maneras particulares en el formalismo clásico, correspondiendo al punto que se encuentra un nivel más abajo en la rama derecha. Ir un nivel inferior de análisis físico requiere el desarrollo de un formalismo nuevo. En este formalismo cuántico, los principios fundamentales son reinstanciados: ocupan el lugar de abajo de la rama izquierda. El formalismo clásico puede ser visto como una descripción de nivel superior de los mismos principios que operan en el nivel cuántico inferior: la línea puntada de la figura 5. Por supuesto, la mecánica cuántica no *implementa* la mecánica clásica: los enfoques están íntimamente relacionados, pero la mecánica clásica provee un enfoque de nivel superior aproximado, no exacto.[4] En un sentido profundo, las teorías cuántica y clásica son totalmente incompatibles: de acuerdo con la ontología de la mecánica cuántica, la ontología de la mecánica clásica es imposible de realizar en este mundo. Pero no se niega que la ontología clásica y los principios que la acompañarían sean teóricamente esenciales, por dos razones al menos: (a) para proporcionar explicaciones (en un sentido literal, explicaciones aproximadas) del enorme rango de fenómenos clásicos para los que una explicación directa a partir de principios cuánticos no tiene ninguna esperanza de ser factible, y (b), históricamente, para proveer la guía necesaria para describir los principios cuánticos, en primer lugar. Tratar de desarrollar principios de nivel inferior sin mirar los principios de nivel superior como guía, dadas las intuiciones que hemos obtenido de esos principios, parecería no aconsejable, para decir lo menos. Básicamente es esta consideración pragmática la que motiva el principio de correspondencia cognitiva y la posición del TAC, a la cual conduce.

En la metodología TAC es esencial estar en condiciones de analizar las propiedades del nivel superior de la computación conexionista para relacionarlas con las propiedades de la computación simbólica: por

4. Muchos casos análogos a la "implementación" se encuentran en física. Las leyes de Newton proveen una "implementación" de las leyes de Kepler; la teoría de Maxwell "implementa" la ley de Coulomb; los principios cuánticos del átomo de hidrógeno "implementan" la fórmula de Balmer.

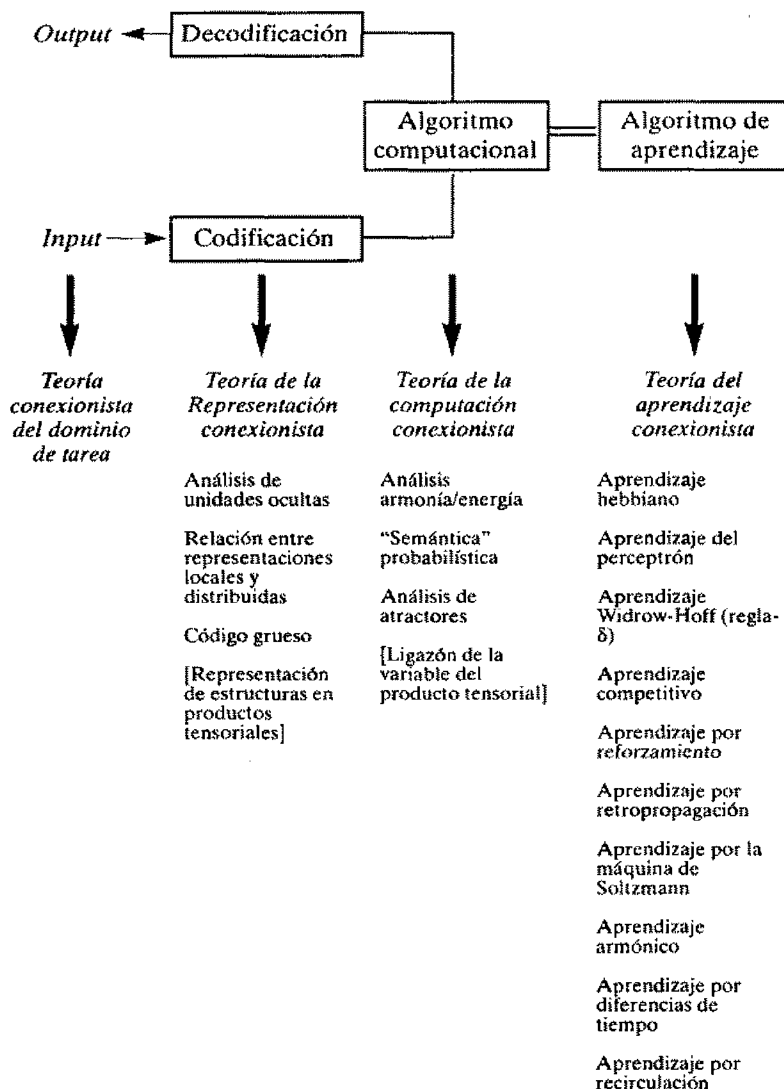


Figura 6. Teoría de los modelos conexionistas

ejemplo, para ver si tienen el poder computacional necesario. Ahora, quiero resumir lo que tomo como el "estado del arte" del análisis matemático de la computación en los sistemas conexionistas, y cómo se relaciona con la crítica de Fodor y Pylyshyn. Este resumen se presenta en la figura 6.

La figura 6 muestra las piezas de un modelo conexionista y los elementos de su análisis. El modelo conexionista tiene básicamente cuatro partes. Está la tarea que se supone que el modelo debe realizar —por ejemplo, tomar un conjunto de *inputs* y conectarlo con un conjunto de *outputs* descriptos en los términos característicos del dominio del problema. Hay luego una red conexionista que ejecutará ese mapeado desde el *input* al *output*; pero entre la primera tarea y el modelo necesitamos métodos para codificar y decodificar. La codificación tiene que tomar la caracterización del *input* en el dominio del problema y codificarlo en una forma tal que la red pueda procesar actividades de ciertos procesadores de *input*. De modo similar, la actividad de los procesadores de *output* tiene que ser decodificada en un enunciado del dominio del problema que pueda ser interpretado como el *output* de la red. El mapeo [*mapping*] *input-output* dentro de la red es el algoritmo computacional encarnado en la red y, bastante a menudo, hay además, un algoritmo de aprendizaje que modifica los parámetros en el algoritmo computacional para lograr que converja en el comportamiento de *input/output* correcto de la computación correcta.

Al analizar estos cuatro elementos de la modelización conexionista, las cosas empeoran de modo progresivo cuando nos movemos de derecha a izquierda. En el área del aprendizaje conexionista, hay muchos análisis: algoritmos para atrapar la fuerza de la conexión del nivel inferior, que producirán una razonable convergencia de nivel superior hacia el mapeo *input/output* correcto. La figura muestra muchos que se ajustan convenientemente, y hay muchos más.[5]

5. He aquí referencias superficiales a estas reglas de aprendizaje; en vez de dar referencias primarias históricas he citado exposiciones recientes fácilmente accesibles que incluyen las citas originales. (De hecho he elegido artículos en Rumelhart, McClelland, y el Grupo PDP, 1986, en lo posible). Para una exposición del aprendizaje hebbiano [*Hebbian learning*], del perceptrón [*perceptron learning*], y del Widrow-Hoff o el aprendizaje por regla-delta [*delta-rule learning*], ver Rumelhart, Hinton y McClelland, 1986, y Stone, 1986. Para el aprendizaje competitivo [*competitive learning*] ver Grossberg, 1987, y Rumelhart y Zipser, 1986. Para el aprendizaje por reforzamiento [*reinforcement learning*], ver Barto, Sutton y Anderson, 1983, y Sutton, 1987. Para aprendizaje por retropropagación [*back propagation learning*] ver Rumelhart, Hinton y Williams, 1986. Para la máquina de aprendizaje Boltzmann [*Boltzmann machine learning*], ver Hinton y Sejnowski, 1986.

Así, si uno piensa que el problema con el conexionismo es que un algoritmo de aprendizaje particular tiene alguna característica que a uno no le gusta, hay posibilidad de que haya otro algoritmo de aprendizaje que lo haga feliz. En cuanto al resto, la teoría del aprendizaje está en buen "estado", aunque cuando llega a los teoremas acerca de qué funciones pueden aprenderse mediante un algoritmo dado, hay muy poco.

Con respecto a analizar las propiedades del nivel superior de los algoritmos para computar *outputs* a partir de *inputs*, hay considerablemente, menos teoría. La técnica de analizar la convergencia usando una función que mide la "energía" o la "armonía" de los estados de la red (Ackley, Hinton y Sejnowski, 1985; Cohen y Grossgerg, 1983; Geman y Geman, 1984; Hinton y Sejnowski, 1983; Hopfield, 1982; Smolensky, 1983, 1986a) nos lleva a alguna parte, como lo hacen otras técnicas,[6] pero parece bastante claro que el estado de análisis de la computación conexionista está considerablemente menos desarrollada que la del aprendizaje conexionista.

Después de esto las cosas se tornan *muy* tenues. ¿Qué decir de la teoría que subyace a la codificación y decodificación, la teoría acerca de cómo tomar las clases de *inputs* y *outputs* que han de ser representadas respecto de los procesos cognitivos y transformarlas en patrones reales de actividad? En términos generales, eso es magia negra: no hay mucho bajo la forma de análisis. Algunos se han esforzado en explorar las representaciones en unidades ocultas (por ejemplo, Hinton, 1986; Rosenberg, 1987), pero hasta ahora veo poca razón para creer que nuestra comprensión de esas representaciones vaya más lejos que comprender un nodo ocasional o unas pocas propiedades estadísticas. Hay unos pocos análisis[7] simples más, pero no nos llevan muy lejos.

En el extremo izquierdo de la figura 6 está la teoría del entorno de tarea [*theory of the task environment*] que surge de una perspectiva cone-

---

Para el aprendizaje armónico [*harmony learning*], ver Smolensky, 1986a. El aprendizaje por diferencias de tiempo [*temporal difference learning*] está relatado en Sutton, 1987. El algoritmo de aprendizaje por recirculación simple [*recirculation learning algorithm*] es discutido en Smolensky, 1987; la idea ha estado bajo exploración por Hinton y McClelland por varios años, y su primer artículo apareció en 1988.

6. Sobre el dar a la computación conexionista una semántica basada en la inferencia estadística, ver Shastri y Feldman, 1985; Smolensky, 1986a; Golden, 1988.

7. Para alguna exploración simple de la relación entre las representaciones locales y distribuidas, ver Smolensky, 1986b. Para algunas observaciones acerca del poder de la técnica representacional distribuida llamada "código grueso" ["*coarse coding*"], ver Hinton, McClelland Rumelhart, 1986.

xionista. Es esencialmente inexistente. Creo que para muchos, es realmente la última meta: la teoría del dominio en términos conexionistas.

Como la figura 6 aclara, aquí hay un punto débil muy importante: la teoría conexionista de la representación. En particular, hasta hace poco no teníamos ideas sistemáticas acerca de cómo representar estructuras complejas. De hecho, fueron Fodor y Pylyshyn quienes realmente me hicieron pensar acerca de esto, y, en última instancia, me convencieron. El resultado fue la técnica del producto tensorial [*tensor product technique*] para generar representaciones totalmente distribuidas de estructuras complejas (Smolensky, 1987). Por esa razón la representación del producto tensorial está dedicada a Fodor y Pylyshyn. Este esquema representacional es una formalización y generalización de las técnicas representacionales que han sido usadas por partes en los modelos conexionistas... La técnica del producto tensorial proporciona un procedimiento sistemático y disciplinado para representar objetos estructurados complejos. Uno puede probar que la representación del producto tensorial tiene un número de bellas propiedades computacionales, desde el punto de vista del procesamiento conexionista. En ese sentido, es apropiado ver la representación del producto tensorial como ocupando la esquina del nivel inferior de la figura 5: proporciona una formalización que es natural para la computación conexionista de la noción no formal de estructura constitutiva y es un candidato posible para jugar un papel en la ciencia cognitiva conexionista, análogo al jugado por la estructura constitutiva de árboles en la ciencia cognitiva simbólica.

La representación del producto tensorial se basa en el uso de la operación del producto tensorial para ejecutar en el mundo vectorial el análogo de ligar una variable y su valor. La figura 6 muestra el punto en el que la ligazón de variables y las representaciones de estructuras *via* el producto tensorial, calzan en el problema global de analizar los modelos cognitivos conexionistas.

Espero que esta última sección haya hecho más plausible mi hipótesis de trabajo, que entre la visión del conexionismo que Fodor y Pylyshyn atacan —negando la importancia de las representaciones estructuradas y los procesos sensibles a la estructura— y la metodología conexionista que propugnan —la implementación de la arquitectura cognitiva clásica— hay un prometedor campo intermedio en el que una investigación productiva y excitante puede ser abordada.

TRADUCCIÓN: Liza Skidelsky.
REVISIÓN TÉCNICA: Eduardo Rabossi.

RECONOCIMIENTOS

Este trabajo ha sido sostenido por las subvenciones NSF IRI-8609599 y ECE-8617947 , y por una subvención del programa de neurociencia computacional de la Fundación Sloan.

REFERENCIAS BIBLIOGRÁFICAS

Ackley, D.H.; Hinton, G.E. y Sejnowski,T. J.: (1985) "A learning algorithm for Boltzmann machines", *Cognitive Science* 9, 147-169.

Ballard, D. y Hayes, P. J.: (1984) "Parallel logical inference", *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, Rochester, Nueva York, junio.

Barto, A.G.; Sutton, R.S.; Anderson, C.W.: (1983) "Neuronlike elements that can solve difficult learning control problems", *IEEE Transactions on Systems, Man and Cybernetics*, SMC-13, 834-846.

Bever, T.G.; Carrithers, C. y Townsend, D.J.: (1987) "A tale of two brains: The sinistral quasimodularity of language", *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, 764-773, Seattle, WA, julio.

Chomsky, N.: (1965) *Aspects of the theory of syntax*, Cambridge, MA., MIT Press.

Cohen, M.A. y Grossberg, S.: (1983) "Absolute stability of global pattern formation and parallel memory storage by competitive neural networks", *IEEE Transactions on Systems, Man and Cybernetics*, SMC-13, 815-826.

Dreyfus, S.E. y Dreyfus, H.L.: (1986) *Mind over machine: The power of human intuition and expertise in the era of the computer*. Nueva York, Free Press.

Fodor, J.A. y Pylyshyn, Z.W.: (1988) "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition*, 28, 2-71.

Feldman, J.A. y Ballard, D.H.: (1982) "Connectionist models and their properties", *Cognitive Science* 6, 205-254.

Geman, S. y Geman, D.: (1984) "Stochastic relaxation. Gibbs distributions and the Bayesian restoration of images". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721-741.

Gerken, L. y Bever, T.G.: (1986) "Linguistics intuitions are the result of interactions between perceptual processes and linguistics universals", *Cognitive Science* 10, 457-476.

Golden, R.: (1988) "An unified framework for connectionist systems", *Biological Cybernetic* .

Greeno, J.G.: (1987) "The cognition connection", *The New York Times*, enero, 4, pág. 28.

Grossberg, S.: (1987) "Competitive learning : From interactive activation to adaptive resonance", *Cognitive Science* 11, 23-63.

Gupta, M.; Ragade, R. y Yager, R. (comps.): (1979) *Advances in fuzzy set theory and applications*, Amsterdam, North Holland.

Hinton, G.E.: (1987) "Learning distributed representations of concepts", *Proceedings of the Eigth Annual Meeting of the Cognitive Science Society*, 1-12.

Hinton, G. E.; McClelland, J. L. y Rumelhart, D. E.: (1986) "Distributed representations", en J.L. McClelland, D.E. Rumelhart y PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 2, Psychological and biological models*, Cambridge, MA, MIT Press/ Bradford Books.

Hinton, G.E. y Sejnowski, T.J.: (1983a) "Analysing cooperative computation", *Proceeding of the Fifth Annual Conference of the Cognitive Science Society*, Rochester, Nueva York.

Hopfield, J.J.: (1982) "Neural networks and physical systems with emergent collective computational abilities", *Proceedings of the National Academy of Sciences*, USA 79, 2554-2558.

Kirsh, D.: (1988) "Paying price for cognition", *The Southern Journal of Philosophy* 26 (supl.).

McClelland, J.L.; Rumelhart, D.E. y PDP Research Group: (1986) *Parallel Distributed Processing: Exploration in the Microstructure of Cognition, Volume 2: Psychological and Biological Models*, Cambridge, MA., MIT Press/ Bradford Books.

Rosenberg, C.R.: (1987) "Revealing the structure of NETtalk's internal representations", *Proceeding of the Ninth Annual Meeting of the Cognitive Science Society*, 537-554, Seattle, WA, julio.

Rumelhart, D.E. Hinton, G.E. y Williams, R.J.: (1986) "Learning internal representation by error propagation", en D.E. Rumelhart; J.L. McClelland y PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations*, Cambridge, MA., MIT Press/ Bradford Books.

Rumelhart, D.E.; Hinton, G.E. y McClelland, J.L.: (1986) "A general framework for parallel distributed processing", en D.E. Rumelhart; J.L. McClelland y PDP Research Group, *Parallel distributed proces-*

sing: Explorations in the microstructure of cognition.Volume 1: Foundations, Cambridge, MA., MIT Press/Bradford Books.

Rumelhart, D.E.; McClelland J.L. y PDP Research Group: (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, Cambridge, MA., MIT Press/Bradford Books.

Rumelhart, D. E. y Zipser, D.: (1986) "Feature discovery by competitive learning", en D. E. Rumelhart, J. L. McClelland y PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*, Cambridge, MA., MIT Press/Bradford Books.

Shastri, L. y Feldman, J. A.: (1985) "Evidential reasoning in semantic networks: A formal theory", *Proceedings of the International Joint Conference on Artificial Intelligence*, Los Angeles, CA.

Shortlife, E. H.: (1976) *Computer-based medical consultations: MYCIN*, Nueva York, American Elsevier.

Smolensky, P.: (1983) "Schema selection and stochastic inference in modular enviroments", *Proceedings of the National Conference on Artificial Intelligence*, Washington, D.C.

Smolensky, P.: (1986a) "Information processing in dynamical systems: Foundations of harmony theory", en D. E. Rumelhart, J. L. McClelland y PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*, Cambridge, MA., MIT Press/ Bradford Books.

Smolensky, P.: (1986b) "Neural and conceptual interpretations of parallel distributed processing models", en D.E. Rumelhart, J. L. McClelland y the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and Biological Models*, Cambridge, MA., MIT Press/Bradford Books.

Smolensky, P.: (1987) "On variable binding and the representation of symbolic structures in connectionist systems", Technical Report CU-CS-355-87, Department of Computer Science. University of Colorado at Boulder, febrero. (Versión revisada publicada en *Artificial Intelligence*.)

Smolensky, P.: (1988a) "The Proper Treatment of Connectionism", *The Behavioral and Brain Sciences* 11 (1): 1-74.

Smolensky, P.: (1988b) "Putting together connectionism-again", *The Behavioral and Brain Sciences*, 11 (1).

Smolensky, P.: (próximo a aparecer) *Lectures on connectionist modeling*, Erlbaum.

Stone, G. O.: (1986) "An analysis of the delta-rule and learning statistical associations", en D. E. Rumelhart, J. L. McClelland y PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*, Cambridge, MA., MIT Press/Bradford Books.

Sutton, R. S.: (1987) "Learning to predict by the methods of temporal differences", Technical Report 87-509, 1, GTE Laboratories, Waltham, MA.

Zadeh, L. A.: (1965) "Fuzzy sets", *Information and Control*, 8, 338-353.

Zadeh, L. A.: (1975) "Fuzzy logic and approximate reasoning", *Synthèse* 30, 407-428.

Zadeh, L. A.: (1983) "Role of fuzzy logic in the management of uncertainty in expert systems", *Fuzzy Sets and Systems*, 11, 199-227.

Capítulo 15

MENTES Y CEREBROS SIN PROGRAMAS *

John Searle

El hiato

El objetivo de este capítulo es presentar una explicación provisional de algunas discusiones que he tenido con filósofos y con gente de otras disciplinas.[1] Quiero comenzar situando los puntos en cuestión en un contexto algo más amplio.

Hay una importante laguna en la vida intelectual del siglo veinte. ('Laguna' es quizás un eufemismo para 'escándalo'.) Confiamos en que podemos dar explicaciones de la conducta humana en términos ordinarios, de sentido común. Así, decimos cosas como 'Ese hombre votó por Ronald Reagan porque pensó que Reagan iba a terminar con la inflación'. Tales comentarios son parte de la psicología de sentido común o de la abuela [*common-sense or grandmother psychology*]. Para darle un nombre elegante podemos llamarla 'psicología intencionalista'['*intentionalistic psychology*']. Suponemos también que por debajo de ese nivel de explicación tiene que haber un nivel de explicación neurofisiológica. Pero, realmente, no sabemos cómo dar explicaciones neurofisiológicas

* "Minds and Brains Without Programs", en *Mindwaves*, Colin Blakemore y Susan Greenfield (comps.), Oxford, Blackwell, 1989, págs. 209-233. Con autorización del autor y de *Basil Blackwell*.

1. Este capítulo está basado en una conferencia dada en Oxford cuando estaba en Inglaterra para grabar las conferencias Reith de 1984 para la BBC, conferencia que fue hecha originalmente para una publicación separada. Hay una superposición considerable entre el material de las conferencias Reith y el material de esta conferencia. Las conferencias han sido publicadas como Minds, Brains and Science (BBC Publications, 1984; Harvard Univ. Press, 1984). Pido disculpas a los que escucharon y leyeron las conferencias Reith por la repetición. Publico este artículo en forma separada, en parte porque Colin Blakemore y Susan Greenfield me convencieron de que sería una contribución útil para este volumen, a pesar de repetir material publicado en otra parte, y en parte porque me da la oportunidad de expandir y explicar varios puntos que se expusieron en las conferencias Reith.

de la conducta humana ordinaria. No sabemos cómo hacer aseveraciones tales como 'El hombre votó por Reagan debido a cierta condición en su tálamo'. Esto nos coloca en una situación intelectual embarazosa. Usamos con razonable confianza la psicología de la abuela en el nivel más elevado y pensamos que tiene que haber una ciencia dura sustentándola en el nivel más bajo, pero no tenemos la más vaga idea de cómo funciona el nivel para explicar los casos específicos de conducta humana normal. Usamos la psicología de la abuela todo el tiempo, pero nos avergüenza llamarla ciencia. Nadie, por ejemplo, tiene la presencia de ánimo como para presentarse a la National Science Foundation a pedir una beca para hacer psicología de la abuela. Sin embargo, no sabemos lo suficiente acerca del nivel más bajo como para hacerlo funcionar. Parece, pues, que tenemos un hiato.

psicología intencionalista
——————————— ← hiato
neurofisiología

Algunos de los grandes esfuerzos intelectuales del siglo veinte han sido intentos de salvar el hiato, de encontrar algo que fuera una ciencia de la conducta humana pero que no fuera psicología de sentido común ni tampoco fuera neurofisiología. Y si uno vive lo suficiente, es interesante mirar atrás y ver los cadáveres de teorías que supuestamente iban a salvar el hiato. Durante mi vida el fracaso más espectacular fue el del conductismo. Pero también viví otros esfuerzos fallidos. Hubo la teoría de juegos y la teoría de la información. No creo que nadie que lea esto sea tan viejo como para no recordar la cibernética, pero en un tiempo se hicieron grandes aseveraciones acerca del futuro de la cibernética. Hubo algo llamado 'estructuralismo', que fue seguido por algo llamado 'post-estructuralismo'. Ahora está la sociobiología, otro candidato para salvar el hiato.

Sin embargo, el principal candidato se llama hoy 'ciencia cognitiva', y con frecuencia se piensa que el programa central de investigación en la ciencia cognitiva es la inteligencia artificial. Hay diferentes escuelas de ciencia cognitiva y de inteligencia artificial, pero la teoría más ambiciosa para salvar el hiato es la que dice que la investigación en psicología cognitiva y en inteligencia artificial ha establecido que la mente es al cerebro como el programa del computador es al *hardware* del computador. La siguiente ecuación es muy común en la literatura: mente/cerebro = programa/*hardware*. Para distinguir este punto de vista de versio-

nes más cautelosas de la inteligencia artificial, lo he llamado 'inteligencia artificial fuerte' ('IA fuerte', para abreviar). De acuerdo con la IA fuerte, un computador adecuadamente programado, con los *inputs* y *outputs* correctos, tiene literalmente una mente en el mismo sentido en que usted y yo la tenemos.

Este punto de vista tiene algunas consecuencias interesantes. Tiene la consecuencia, por ejemplo, de que no hay nada esencialmente biológico respecto de la mente humana. Sucede que los programas que son constitutivos de las mentes son operados [*are run*] en el *wetware* que tenemos en nuestra máquina biológica, en el computador biológico de la cabeza. Pero esos mismos programas podrían ser operados en el *hardware* de cualquier computador que fuera capaz de sostener el programa. Y esto tiene la consecuencia adicional de que cualquier cosa, cualquier sistema, podría tener pensamientos y sentimientos —y no sólo *podría tener*, sino que *tendría que tener* pensamientos y sentimientos— exactamente en el mismo sentido en que nosotros los tenemos, con la sola condición de que se esté operando el programa correcto. Esto es, si se tiene el programa correcto, con los *inputs* y *outputs* correctos, entonces cualquier sistema que opere ese programa, al margen de su estructura química (sea que esté hecho de latas de cerveza viejas o de chips de siliconas o de cualquiera otra substancia) *tiene que tener* pensamientos y sentimientos, exactamente de la misma manera en que usted y yo los tenemos. Y esto es así porque en eso consiste tener una mente: en tener el programa correcto. Ahora bien, siempre que ataco este punto de vista, muchos dicen: 'Pero seguramente nadie puede creer eso'. Voy a decirles los nombres de algunas personas que creen eso, así no piensan que estoy atacando a un personaje imaginario.

Herbert Simon, de la Carnegie-Mellon University, ha escrito en numerosas ocasiones que ya contamos con máquinas que pueden pensar en un sentido literal; que pueden pensar en el mismo sentido que usted y yo lo hacemos. Los filósofos han discutido durante siglos si se puede o no construir una máquina que piense, y ahora lo hacen a diario en Carnegie-Mellon. Allan Newell, el colega de Simon, en una conferencia que le escuché dar en San Diego en la reunión fundacional de la Cognitive Science Society, dijo que se había descubierto (no que era alguna hipótesis que se estaba considerando, sino que se había 'descubierto') que la inteligencia es exclusivamente manipulación de símbolos físicos. De modo que cualquier máquina que sea capaz de manipular los símbolos correctos de manera correcta, tiene procesos inteligentes, exactamente en el mismo sentido en que usted y yo los tenemos. Marvin

Minsky dice que la próxima generación de computadores va a ser tan inteligente que vamos a tener suerte si nos dejan en casa como mascotas. Y Freeman Dyson es citado en el *New York Times* como habiendo dicho que dado que ahora sabemos que procesos mentales tales como la conciencia son procesos puramente formales, hay una ventaja evolutiva en tener tales procesos formales (conciencia y demás) funcionando en chips de siliconas y alambres, porque en un universo que se está enfriando, ese tipo de material tiene más capacidad para sobrevivir que los organismos como los nuestros, hechos de una desaliñada maquinaria biológica. Así que el próximo paso evolutivo, según este punto de vista, va a estar hecho de alambres y siliconas. En esta literatura mi texto favorito (y les recomiendo esta literatura porque es maravillosa) es de John McCarthy, el inventor del término 'inteligencia artificial'. McCarthy escribió: "Puede decirse que máquinas tan simples como los termostatos tienen creencias...". Y agregó, por cierto: "Tener creencias parece ser una característica de la mayoría de las máquinas capaces de resolver problemas".[2] De modo que le pregunté: "John, ¿qué creencias tiene tu termostato?". Admiro su coraje. Dijo: "Mi termostato tiene tres creencias. Mi termostato cree que hace demasiado calor aquí, que hace demasiado frío aquí y que la temperatura es adecuada aquí".

Bien, me gusta esta tesis por una simple razón. La ecuación mente/cerebro = programa/*hardware*, no es usual en filosofía porque es una tesis razonablemente clara. Uno la puede enunciar con razonable precisión. Y a diferencia de la mayoría de las tesis filosóficas, está sujeta a una muy simple y, pienso, decisiva refutación. He publicado la refutación en otra parte, pero la repetiré brevemente aquí porque en la comunidad de la inteligencia artificial no se acepta universalmente que haya refutado, de hecho, ese punto de vista. Después quiero tocar una cuestión más profunda, que es ésta: una de las razones por las que la gente cree en la IA fuerte es que no puede ver otra manera de resolver el problema mente-cuerpo. Estoy convencido de que una de las fuentes de la creencia de que tener una mente equivale a tener un programa de computación, es que la gente no puede ver otra forma de resolver el problema mente-cuerpo sin recurrir al dualismo. Con frecuencia se me pregunta: "Bien, si no acepta el análisis de la mente que ofrece la inteligencia artificial, entonces, ¿cuál es su solución al problema mente-cuer-

2. McCarthy, John (1979): "Ascribing mental qualities to machines", Stanford Artificial Intelligence Laboratory Memo AIM-326, pág. 2, *Computer Science Department Report*, No. STAN-CS-79-725, marzo de 1979.

po? ¿No está usted forzado a caer en el dualismo o en el misticismo o en el vitalismo o en algún enfoque, igualmente misterioso?". Así que realmente tengo dos tareas. Quiero refutar a la inteligencia artificial fuerte y quiero resolver el problema mente-cuerpo.

Una nueva visita a la habitación china

El argumento en contra de la IA fuerte es, me temo, bastante simple. El argumento se me ocurrió cuando leí el libro de Schank y Abelson acerca de sus programas de comprensión-de-relatos [*story-understanding programs*].[3] Algunos estarán familiarizados con ellos; pero volveré a repasar cómo funcionan sus programas. Se trata de programas muy ingeniosos que han sido diseñados en Yale University. Los programas realizan lo que ellos llaman 'comprender relatos'. El computador recibe un relato muy simple como *input*. Una historia típica sería la que sigue:

'Un hombre fue a un restaurante y pidió una hamburguesa. Cuando le trajeron la hamburguesa estaba muy quemada. El hombre salió enfurecido del restaurante sin haber pagado la hamburguesa.'

Entonces uno le pregunta al computador: '¿Comió el hombre la hamburguesa?'. Y el computador dice como *output*: 'No, el hombre no comió la hamburguesa'. O uno le da a la computadora otro relato:

'Un hombre fue a un restaurante y pidió una hamburguesa. Cuando le trajeron la hamburguesa se deleitó con ella, y cuando dejó el restaurante pagó la cuenta y dejó una buena propina al mozo'.

Si entonces uno le pregunta al computador: '¿Comió el hombre la hamburguesa?'. El computador responde: 'Sí, el hombre comió la hamburguesa'. Nótese que ninguna de las dos historias decía explícitamente si el hombre había comido o no la hamburguesa. ¿Cómo funciona? Funciona porque el programa tiene en su base de datos lo que se llama un 'guión-de-restaurante' [*restaurant script*]. El guión-de-restaurante es la representación de cómo son las cosas normalmente en los restaurantes. Cuando el computador toma el relato, lo aparea con el guión-de-restaurante, y entonces, cuando toma la pregunta acerca del relato, aparea la pregunta con ambos: el relato y el guión-de-restaurante. Dado que 'sabe' cómo se supone que suceden las cosas en los restaurantes, puede producir la respuesta adecuada. La alegación, que se hace con frecuen-

3. Schank, R.C. y Abelson, R.P. (1977): *Scripts, Plans, Goals and Understanding*, Hillsdale, NJ., Erlbaum.

cia respecto de los programas es que como la máquina satisface el test de Turing, la máquina tiene que comprender literalmente el relato.[4] Tiene que comprender literalmente el relato en el mismo exacto sentido en que usted y yo comprenderíamos esos relatos si nos hicieran tales preguntas y diéramos buenas respuestas.

Me parece que hay una refutación muy simple a esa alegación. La refutación consiste en imaginar que uno es la máquina. A mí me gusta imaginarla de la siguiente manera.

Supóngase que estoy encerrado en una habitación. En esa habitación hay dos grandes cestos llenos de símbolos chinos, junto con un libro de reglas en español acerca de cómo aparear los símbolos chinos de una de las cestas con los símbolos chinos de la otra cesta. Las reglas dicen cosas como: 'Busque en la canasta 1 y saque un signo garabateado, y póngalo al lado de un cierto signo garabateado que saque de la canasta 2'. Adelantándonos un poco, esto se llama una 'regla computacional definida en base a elementos puramente formales'. Ahora bien, supóngase que la gente que está fuera de la habitación envía más símbolos chinos junto con más reglas para manipular y aparear los símbolos. Pero esta vez sólo me dan reglas para que les devuelva los símbolos chinos. Así que estoy aquí, en mi habitación china, manipulando estos símbolos. Entran símbolos y yo devuelvo los símbolos de acuerdo con el libro de reglas. Ahora bien, sin yo saberlo, quienes organizan todo esto fuera de la habitación llaman a la primera cesta un 'guión-de-restaurante' y a la segunda cesta un 'relato acerca del restaurante', a la tercera hornada de símbolos la llaman 'preguntas acerca del relato,' y a los símbolos que les devuelvo 'respuestas a las preguntas'. Al libro de reglas lo llaman 'el programa', ellos se llaman 'los programadores' y a mí me llaman 'el computador'. Supóngase que después de un tiempo soy tan bueno para responder esas preguntas en chino que mis respuestas son indistinguibles de las de los nativos hablantes del chino. Con todo, hay un punto muy importante que necesita ser enfatizado. Yo no comprendo una palabra del chino, y no hay forma de que pueda llegar a entender el chino a partir de la instanciación de un programa de computación, en la manera en que la describí. Y éste es el quid del relato: *si yo no comprendo chino en esa situación, entonces tampoco lo comprende ningún otro computador digital sólo en virtud de haber sido adecuadamente programado, porque ningún computador digital por el solo hecho de ser un computador digital, tiene algo que yo no tenga.* Todo lo que tiene un

4. No quiero implicar que Schank o Abelson hagan esta afirmación.

computador digital, por definición, es la instanciación de un programa formal de computación. Pero como yo estoy instanciando el programa, como suponemos que tenemos el programa correcto con los *inputs* y *outputs* correctos, y yo no comprendo el chino, entonces no hay forma de que cualquier otro computador digital *sólo en virtud de instanciar el programa* pueda comprender el chino.

Éste es el núcleo central del argumento. Pero su quid, pienso, se perdió en el montón de bibliografía desarrollada subsiguientemente a su alrededor; así que quiero enfatizarlo. El quid del argumento no es que de una u otra manera tenemos la 'intuición' de que no comprendo el chino, de que me *inclino a decir* que no comprendo el chino pero que, quien sabe, quizá realmente lo entienda. Ese no es el punto. El quid del relato es recordarnos una verdad conceptual que ya conocíamos, a saber, que hay diferencia entre manipular los elementos sintácticos de los lenguajes y realmente comprender el lenguaje en un nivel semántico. Lo que se pierde en la *simulación del* comportamiento cognitivo de la IA, es la distinción entre la sintaxis y la semántica.

El quid del relato puede enunciarse ahora más genéricamente. Un programa de computación, por definición, tiene que ser definido de manera puramente sintáctica. Es definido en términos de ciertas operaciones formales realizadas por la máquina.[5] Eso es lo que hace del computador digital un instrumento tan poderoso. Uno y el mismo sistema de *hardware* puede instanciar un numero indefinido de programas de computación diferentes, y uno y el mismo programa de computación puede operarse en *hardwares* diferentes, porque el programa tiene que ser definido de manera puramente formal. Pero por esa razón la simulación formal de la comprensión del lenguaje nunca va a ser en sí lo mismo que la duplicación. ¿Por qué? Porque en el caso de comprender realmente un lenguaje, tenemos algo más que un nivel formal o sintáctico. Tenemos la semántica. No manipulamos meramente símbolos formales no interpretados, sabemos realmente qué significan.

Esto puede mostrarse enriqueciendo un poco el argumento. Estoy allí, en la habitación china, manipulando esos símbolos chinos. Supóngase ahora que algunas veces los programadores me dan relatos en español y que me hacen preguntas, también en castellano, acerca de esos

5. Los programas de 'Comprensión lingüística' ['*Language understanding*'] tienen característicamente una 'sintaxis', una 'semántica' y en algunos casos hasta una 'pragmática'. Por supuesto, esto es irrelevante para el argumento, porque los tres niveles son computacionales, esto es, 'sintácticos', en el sentido en que estoy usando ahora la palabra.

relatos. ¿Cuál es la diferencia entre los dos casos? Tanto en el caso del español, como en el caso del chino, satisfago el test de Turing. Esto es, doy respuestas que son indistinguibles de las respuestas que daría un hablante nativo. En el caso del chino lo hago porque los programadores son buenos para diseñar el programa, y en el caso del español porque soy un hablante nativo. ¿Cuál es la diferencia, entonces, si mi actuación es equivalente en los dos casos? Me parece que la respuesta es obvia. La diferencia es que sé español. Sé qué significan las palabras. En el caso del español no sólo tengo una sintaxis, tengo una semántica. Atribuyo un contenido semántico o significado a cada una de esas palabras, y por lo tanto estoy haciendo más que lo que un computador digital puede hacer en virtud de instanciar un programa. Tengo una interpretación de las palabras y no sólo de los símbolos formales. Nótese que si tratamos de dar al computador una interpretación de los símbolos formales lo único que podemos hacer es darle más símbolos formales. Todo lo que podemos hacer es poner más símbolos formales no interpretados. Por definición, el programa es sintáctico, y la sintaxis por sí misma nunca es suficiente para la semántica.[6]

Bueno, en esto consiste mi rechazo de la ecuación, mente/cerebro = programa/*hardware*. Instanciar el programa correcto nunca es suficiente para tener una mente. Tener una mente es algo más que instanciar un programa de computación. Y la razón es obvia. Las mentes tienen contenidos mentales. Tienen contenidos semánticos Así como tienen un nivel sintáctico de descripción.

Hay un mal entendido persistente acerca de mi argumento, que quiero neutralizar ya. Algunos suponen que sostengo que en principio es imposible para los chips de siliconas duplicar el poder causal del cerebro. Ése no es mi argumento; es más, no tiene ninguna conexión con mi argumento. Que los poderes causales de las neuronas puedan ser duplicados en algún otro material, como chips de siliconas, válvulas electrónicas, transistores, latas de cerveza, o alguna desconocida substancia química, es una cuestión fáctica, que no ha de ser resuelta apelando a bases puramente filosóficas o a priori. El quid de mi argumento es que uno no puede duplicar los poderes causales del cerebro sólo en virtud

6. Alguna gente verdaderamente temeraria de la IA ha propuesto que no soy yo quien comprende sino la habitación entera, esto es, el sistema que me contiene, el programa, los canastos, la ventana al exterior, etcétera. Pero esta respuesta está sujeta a la misma objeción. Así como yo no tengo manera de pasar de la sintaxis a la semántica, tampoco lo tiene el sistema total. El sistema total no tiene manera de saber qué significa cualquiera de los símbolos formales.

de instanciar un programa de computación, debido a que el programa de computación tiene que ser definido de manera puramente formal. Es importante enfatizar que la inteligencia artificial, sea fuerte o de otra manera, no tiene nada que ver con las propiedades químicas de la silicona o de cualquiera otra sustancia. Una vez que el partidario de la IA concede que tales propiedades son relevantes, ha abandonado la tesis de la IA. La IA es acerca de los poderes 'cognitivos' de los programas. Nada tiene que ver con las propiedades químicas específicas de las realizaciones de los programas en el *hardware*.

Sin embargo, esto nos enfrenta a la segunda cuestión. Si rechazamos la ecuación y rechazamos la IA como salvando el hiato, entonces, ¿cuál es nuestro análisis de la relación entre el nivel de la intencionalidad y el nivel de la neurofisiología? Una respuesta breve es que la razón de que lo que cierra el hiato siempre falle, es que no hay ningún hiato que salvar. No hay hiato alguno entre el nivel de las explicaciones intencionalistas y el nivel de las explicaciones neurofisiológicas. Pero para fundamentar esto necesito, como prometí antes, resolver el problema mente-cuerpo.

Cuatro enigmas

Antes de considerar directamente el problema mente-cuerpo, quiero volver atrás por un momento y preguntar por qué este problema parece ser tan dificultoso. ¿Por qué en filosofía, psicología y neurofisiología, todavía tenemos el problema mente-cuerpo? Desde Descartes, al menos, la forma general del problema mente-cuerpo ha sido el problema de reconciliar nuestras creencias de sentido común y precientíficas acerca de la mente con nuestra concepción científica de la realidad. Nuestra concepción científica del mundo, como un sistema físico o como un conjunto de sistemas físicos en interacción, ha crecido en poder y comprehensividad, y parece cada vez más difícil encontrar en esa concepción un lugar para la mente. Algunos de los puntos de vista precientíficos que parecen ser cuestionados por el crecimiento de una visión científica del mundo derivan de la religión o de la moral —doctrinas tales como la inmortalidad del alma, el libre albedrío, la naturaleza de la responsabilidad moral—, y acerca de esas cuestiones no tendré nada que decir en esta discusión. Me detendré en una pregunta más restringida y, creo, más apremiante: ¿cómo podemos encuadrar lo que sabemos o lo que creemos saber acerca del mundo en general, con lo que

sabemos o lo que creemos saber acerca del funcionamiento de nuestras propias mentes? Dejando a un lado las especulaciones de la religión y las presuposiciones de la moral, sabemos una cantidad de cosas acerca de nuestras mentes, y mi objetivo es dar una explicación coherente de las relaciones entre lo que sabemos acerca de nuestras propias mentes y lo que sabemos acerca de la forma en que funciona el mundo en general. ¿Por qué este problema más restringido, no religioso y no moral, ha sido tan dificultoso? ¿Por qué, para decirlo una vez más, hay todavía un problema mente-cerebro o mente-cuerpo?

Las características [*features*] de nuestra concepción de sentido común de la mente que parecen difíciles de asimilar a nuestra concepción científica general del mundo son, al menos, las cuatro siguientes:

Conciencia

Yo, en el momento de escribir esto, y usted, en el momento de comprenderlo, somos concientes. Que el mundo contiene estados mentales concientes es un hecho evidente, pero es difícil ver cómo meros sistemas físicos pueden tener conciencia. ¿Cómo puede ocurrir tal cosa? ¿Cómo, por ejemplo, puede ser conciente este trozo de materia gris y blanca que está adentro de mi cráneo?

Intencionalidad

Muchos de mis estados mentales, como por ejemplo, mis creencias y deseos y mis percepciones visuales e intenciones, están dirigidas a, o son acerca de objetos y estados en el mundo, distintos de ellos mismos. Este rasgo, llamado 'intencionalidad' es una característica de las mentes humanas. Pero, nuevamente, ¿cómo puede ocurrir una cosa tal? ¿Cómo pueden ser *acerca* de algo procesos en mi cerebro que, después de todo, consisten finalmente en 'átomos en el vacío' ['*atoms in the void*']? ¿Cómo pueden átomos en el vacío *representar* algo? Uno se inclina a decir: las cosas y los procesos en el mundo simplemente son; sea que pensemos en *procesos* como la digestión y en secuencias de neuronas excitadas o en *cosas* físicas corrientes como piedras y árboles, parece imposible que alguna de ellas pueda ser *acerca de* algo. ¿Cómo puede el *acerca de algo* [*aboutness*] ser un rasgo intrínseco del mundo?

Subjetividad

Los estados mentales son característicamente subjetivos. Pero es difícil entender cómo el mundo físico objetivo, igualmente abierto a todos los observadores competentes, puede contener algo esencialmente subjetivo como, por ejemplo, estados mentales concientes. Interpretada ingenuamente, la subjetividad de los estados mentales está marcada por hechos tales como que yo tengo mis estados y no los suyos, que mis estados me son accesibles de una manera en que no son accesibles para usted; percibo el mundo desde mi punto de vista y no desde su punto de vista, etcétera. ¿Cómo puede la subjetividad ser una parte real del mundo?

Causación intencional

Aun si hubiera cosas tales como estados mentales, es difícil ver cómo ellos podrían producir una diferencia real en el mundo. ¿Podría algo, por decirlo de alguna manera, tan 'gaseoso' y 'etéreo' como un estado mental conciente tener algún impacto en un objeto físico como el cuerpo humano? ¿Cómo podrían los fenómenos mentales empujar objetos o tener cualquiera otra significación física? ¿No serían los eventos mentales, si existieran, sólo epifenoménicos?

Llamemos a estos problemas, respectivamente, los problemas de la conciencia, de la intencionalidad, de la subjetividad y de la causación intencional. Aunque no todos los estados mentales tienen estas cuatro características, son sin embargo características reales y típicas de los fenómenos mentales. Sabemos, por ejemplo, que las personas están a menudo en un estado de conciencia, que tienen con frecuencia, por ejemplo, pensamientos y sentimientos que refieren a objetos y estados de cosas fuera de ellas mismas, que aprehenden el mundo desde un punto de vista subjetivo, y que sus pensamientos y sentimientos tienen relevancia en su comportamiento. Creo que cualquier explicación acerca del problema mente-cerebro debe ser capaz, al menos, de dar cuenta de todos estos hechos.

En el punto de vista que se adopta en este ensayo acerca de los estados mentales, ellos y los procesos mentales son fenómenos biológicos reales en el mundo, tan reales como la digestión, la fotosíntesis, la lactancia o la secreción de bilis. El objetivo de este capítulo no es mostrar en detalle cómo tales fenómenos biológicos están relacionados con los procesos neurofisiológicos del cerebro —nadie sabe en detalle cómo

están relacionados—, su objetivo, en cambio, es más modesto: mostrar cómo es posible que los estados mentales puedan ser fenómenos biológicos en el cerebro. Creo que un supuesto típico pero no enunciado de muchas de las implausibles doctrinas contemporáneas acerca de la mente —doctrinas tales como el conductismo o la inteligencia artificial fuerte— es que es sencillamente imposible acomodar una explicación ingenua de sentido común acerca de la mente con una visión científica del mundo. Y creo que es la desesperación causada por el sentimiento de que no puede darse ninguna explicación coherente que acomode el mentalismo de sentido común con la ciencia dura, lo que ha llevado a la gente a decir las cosas implausibles y algunas veces tontas que se dicen acerca de la naturaleza de la mente. El punto de vista que voy a exponer acerca de la relación de la mente y el cerebro, es consistente con lo que se sabe acerca del funcionamiento del cerebro y es también consistente con un enfoque biológico general de los fenómenos biológicos. Mi enfoque no intenta tratar a la mente como algo formal o abstracto, tal como hace la IA fuerte, ni tampoco intenta tratar a la mente simplemente como un conjunto neutral de poderes causales sin características mentales intrínsecas, tal como hacen ciertas formas de funcionalismo. Francamente pienso que el enfoque que voy a presentar es más bien un punto de vista obvio y de sentido común, y hasta que me vi envuelto en esas polémicas recientes, suponía que era ampliamente aceptado, tan ampliamente aceptado que hasta no merecía una enunciación expresa. Sin embargo, mis formulaciones previas han sido calificadas por mis críticos de 'místicas' (Ringle),[7] 'sofísticas' (Dennett),[8] 'religiosas' (Hofstadter),[9] etcétera. Quizás, entonces, valga la pena explicar la posición con algún detalle para que se pueda ver que esos cargos son realmente infundados. No preciso enfatizar que no soy la primera persona en sostener ese punto de vista y que similares enfoques biológicos del problema mente-cuerpo pueden encontrarse al menos tan atrás como el siglo diecinueve.

7. Ringle, Martin (1980): "Mysticism as a philosophy of artificial intelligence", comentario sobre "Minds, brains and programs" de Searle, *The Behavioral and Brain Sciences*, 3: 444.

8. Dennett, Daniel (1980): "The milk of human intentionality", comentario sobre "Minds, brains and programs" de Searle, *The Behavioral and Brain Sciences*, 3: 428.

9. Hofstadter, Douglas R. (1980): "Reductionism and religion", comentario sobre "Minds, brains and programs", *The Behavioral and Brain Sciences*, 3: 433.

El cerebro y su mente

¿Cómo funciona el cerebro? En detalle, nadie lo sabe. Yo tengo una ignorancia de *amateur* respecto de este tema, pero aun los mejores expertos están confundidos, hasta ahora, respecto de lo que uno pensaría que son las preguntas más fundamentales. ¿Cuál es exactamente la neurofisiología de la conciencia? ¿Por qué necesitamos dormir? ¿Cómo se almacenan los recuerdos en el cerebro, con exactitud? ¿Por qué el alcohol nos emborracha? ¿Por qué la aspirina alivia el dolor? Recientemente, en 1978, un neurólogo famoso, David Hubel, escribió: 'Hay [áreas del cerebro] del tamaño de un puño, de las cuales se puede decir que estamos casi en el mismo estado de conocimiento del que estábamos con relación al corazón antes de darnos cuenta de que bombeaba sangre'.[10] Más aún, en nuestra ignorancia, buscamos a tientas metáforas y analogías, generalmente basadas en la última tecnología. Así, hoy en día, el punto de vista de moda es que el cerebro es un computador digital, pero en mi niñez se aseguraba que era un tipo de tablero de distribución telefónico; Charles Sherrington comparó al cerebro con un sistema de telégrafo y con un telar de *jacquard*; Sigmund Freud lo comparó con bombas hidráulicas y sistemas electromagnéticos; Leibniz con un molino, y me han dicho que ciertos griegos antiguos pensaban que el cerebro funcionaba como una catapulta. El último punto de vista entre los neurofisiólogos es que el cerebro funciona como un sistema de selección natural darwiniano.

Sin embargo, aunque haya mucho para aprender, no somos totalmente ignorantes, y en una discusión como ésta necesitamos recordar unas pocas cosas elementales acerca del cerebro. Como todos los órganos, el cerebro consiste en células. Sin embargo, a diferencia de los demás órganos, el cerebro y el resto del sistema nervioso consiste, en gran parte, en un tipo muy especial de células: las neuronas. Las estimaciones corrientes dicen que hay entre 50 y 100 mil millones de neuronas en el cerebro humano. Hay una gran diversidad de tipos de neuronas, pero la neurona típica consiste en un cuerpo celular o soma, con dos tipos de fibras largas que emergen de él, un único axón y una cantidad de dendritas. Las neuronas se ponen en contacto unas con otras en ciertas protuberancias pequeñas llamadas sinapsis. Los axones y las dendritas realmente no se funden en las sinapsis; el axón tiene, como caracte-

10. Hubel, D. (1978): "Vision and the brain", *Bulletin of the American Academy of Arts and Sciences*, abril de 1978, 31, No. 7, 18.

rística, una pequeña protuberancia, el botón [*bouton*], que toca en la punta la dendrita, y la pequeña brecha que queda entre ellos es el espacio sináptico [*synaptic cleft*]. También hay sinapsis en el soma. Algunas neuronas en el cerebelo tienen hasta 200.000 sinapsis en una célula. Una de las funciones básicas de la neurona es la transmisión de impulsos eléctricos, esto es, cambios 'todo-o-nada' en el potencial eléctrico. Cada impulso eléctrico pasa del soma a lo largo del axón. Sin embargo, en la mayoría de las neuronas el impulso eléctrico no pasa directamente de una neurona a la siguiente; más bien, el impulso eléctrico, cuando llega al botón, causa la liberación de pequeñas cantidades de fluido de los pequeños compartimientos del botón, las vesículas sinápticas, al espacio sináptico. La liberación de esos fluidos (los neurotransmisores) en las sinapsis, puede tener un efecto excitatorio o inhibitorio en la neurona siguiente. Si es excitatorio, va a tender a causar el disparo [*firing*] de la neurona siguiente o va a incrementar el índice de disparo [*rate of firing*]. Si es inhibitorio, va a tender a impedir que la neurona se dispare o a disminuir el índice de disparo. Desde un punto de vista funcional lo importante no es que la neurona dispare, porque de todas maneras muchas neuronas disparan permanentemente. Lo que es importante son las variaciones del *índice* de disparo de las neuronas; específicamente, las variaciones en el índice de disparos de los axones respecto de la suma de todas las excitaciones e inhibiciones en las dendritas.

Es importante enfatizar este punto porque muchos autores han supuesto, erróneamente, que el carácter 'todo-o-nada' del disparo de los impulsos nerviosos constituye una prueba de que los principios del funcionamiento del cerebro son los de un computador digital.[11] Nada puede estar más alejado de la verdad. Hasta donde sabemos, el aspecto funcional de la neurona es la variación no-digital en el índice de disparo.

En la descripción tradicional del cerebro, es decir, la descripción que toma a la neurona como la unidad fundamental de funcionamiento del cerebro, lo más importante acerca de la relación entre el cerebro y la mente es simplemente esto. Toda la enorme variedad de *inputs* que recibe el cerebro —los fotones que alcanzan la retina, las ondas sonoras que estimulan las células del oído interno, la presión en la piel que activa las terminales nerviosas correspondientes a presión, calor, frío y dolor, etcé-

11. Oppenheim, Paul y Putnam, Hilary (1958): "Unity of science as a working hypothesis", en Feigl, Scriven y Maxwell (comps.), *Minnesota Studies in the Philosophy of Science*, vol. 2, *Concepts, Theories, and the Mind-Body Problem*, pág. 19, Minneapolis, Univ. of Minnesota Press.

tera— todos estos *inputs* son convertidos a un medio común: índices variables de disparos de neuronas. Más aún, y de manera igualmente notable, esos índices variables de disparos de las neuronas relativos a diferentes circuitos neuronales y a diferentes condiciones locales en el cerebro, producen toda la variedad y heterogeneidad de la vida mental del agente humano o animal. El aroma de una rosa, la experiencia del azul del cielo, el gusto de las cebollas, el pensamiento de una fórmula matemática, todo esto es producido por índices variables de disparos de neuronas, en circuitos diferentes relativos a condiciones locales diferentes del cerebro. Ahora bien, ¿qué son exactamente esos circuitos neuronales diferentes y cuáles son los entornos locales diferentes que dan cuenta de las diferencias en nuestra vida mental? En detalle nadie lo sabe, pero tenemos buena prueba de que ciertas regiones del cerebro se especializan en cierto tipo de experiencias. El córtex visual juega un rol especial en las experiencias visuales, el córtex auditivo en experiencias auditivas, etcétera. La visión es una de las funciones del cerebro mejor comprendida (o menos inadecuadamente comprendida), y en el caso de la visión parece haber neuronas muy especializadas en el córtex visual capaces de responder a diferentes rasgos específicos de los estímulos visuales. Supóngase que estímulos auditivos alimentaran al córtex visual y que estímulos visuales alimentaran al córtex auditivo. ¿Qué sucedería? Hasta donde sé, nadie ha realizado jamás ese experimento, pero parece razonable suponer que el estímulo auditivo sería 'visto', esto es, que produciría experiencias visuales, y que los estímulos visuales serían 'escuchados', esto es, producirían experiencias auditivas, debido en ambos casos a rasgos específicos aunque ampliamente desconocidos del córtex visual y del auditivo, respectivamente. Aunque esta hipótesis es especulativa, tiene algún soporte independiente si se piensa que un golpe en el ojo produce un *flash* visual ('ver las estrellas'), aun cuando no sea un estímulo óptico.

En mi visión de lego la cantidad de conocimiento que tenemos actualmente acerca de la naturaleza y el funcionamiento de las neuronas, es bastante impresionante: sin embargo, existe ahora fuerte evidencia de que para entender el papel del cerebro en la vida mental necesitamos entender el funcionamiento del cerebro en niveles más elevados que el de las neuronas individuales, y, en particular, que en esos niveles elevados necesitamos entender el funcionamiento de los sistemas de neuronas organizadas en redes neurales o en circuitos neurales. Para muchas funciones del cerebro la unidad de funcionamiento no es la célula singular sino la red de células y, en ese nivel, el funcionamiento del

cerebro consiste en la interacción entre un gran conjunto de redes neurales. La mejor prueba anatómica en favor de la existencia de redes que funcionan más o menos independientemente, es la existencia de módulos neurales en la forma de columnas verticalmente orientadas, de cilindros o de lajas [*slabs*] de células en el córtex.[12] Como dice Gerald Edelman, 'El mayor logro en el pensamiento acerca del córtex, la mayor revolución, es que el córtex no es una hoja [*sheet*] continua dispuesta horizontalmente, sino que está verticalmente organizado como pilas de lajas o columnas'.[13] Estos módulos pueden variar en la cantidad de neuronas que contienen, desde 50 hasta 50.000 o aún más. Desde el punto de vista modular, la significación actual de la neurona reside en la contribución que hace al funcionamiento del módulo.

No sé si la combinación de la explicación neuronal y de la explicación modular o quizá de algún otro tipo de explicación, es la explicación correcta del funcionamiento del cerebro. Pero una conclusión surge con claridad aun de la más corriente investigación del funcionamiento del cerebro: *los fenómenos mentales, concientes o inconcientes, visuales o auditivos, los dolores, cosquilleos, picazones, pensamientos, y el resto de nuestra vida mental, son causados por procesos que suceden en el cerebro.* Los fenómenos mentales son un resultado de los procesos electroquímicos en el cerebro, tanto como la digestión es el resultado de procesos químicos que suceden en el estómago y en el resto del aparato digestivo. Creo que éste es un hecho obvio acerca de cómo funciona el mundo y sin embargo sus implicaciones completas generalmente no son percibidas por los estudiosos de la inteligencia artificial, la ciencia cognitiva o la filosofía. También es importante enfatizar que los procesos causales relevantes son enteramente internos al cerebro. Aunque *de hecho* los eventos mentales median entre los estímulos externos y las respuestas motoras, no hay una conexión *esencial*. Un hombre puede, por ejemplo, tener un dolor terrible sin tener un estímulo de dolor [*pain stimulus*] en los nervios periféricos o un comportamiento correspondiente a dolor [*pain behavior*]. Este simple hecho es suficiente para desacreditar toda la tradición conductista en filosofía.

Para substanciar esto, un poco al menos, contemos una parte del

12. Véase J. Szentagothai, "The Brain-Mind Relation: A Pseudoproblem?", en la compilación a la que pertenece este trabajo.

13. Edelman, Gerald (1982): "Through a computer darkly: group selection and higher brain function", *Bulletin of the American Academy of Arts and Sciences*, octubre de 1982, 36, No. 1, 28.

relato causal que corresponde a un tipo de fenómeno mental conciente: el dolor. Las señales de dolor son transmitidas desde las terminales sensoriales nerviosas a la médula espinal por dos tipos de fibras: las fibras A delta se especializan en sensaciones de picazón [*prickling*] y las fibras C se especializan en sensaciones de ardor y malestar [*burning and aching sensations*]. En la médula espinal pasan a través del tracto de Lissauer y terminan en las neuronas de la médula. A medida que las señales suben a la espina dorsal y entran en el cerebro se separan en dos rutas: la ruta de picazón y la ruta de ardor. Ambas rutas pasan a través de una estructura llamada tálamo, pero, más allá de ese nivel, el dolor de picazón [*prickling pain*] se localiza más claramente en el córtex sensorial somático, específicamente en el área somática 1, mientras que la ruta del dolor de ardor [*burning pain*] transmite señales no sólo hacia arriba en el córtex, sino también lateralmente en el hipotálamo y en otras regiones basales del cerebro. Como consecuencia de estas diferencias es mucho más fácil localizar una sensación de picazón; mientras que las sensaciones de ardor y malestar pueden ser más dificultosas porque activan más 'partes' del sistema nervioso. La sensación real de dolor parece ser causada por la estimulación de las regiones basales, especialmente el tálamo, y por la estimulación del córtex sensorial somático.

Ahora bien, a los fines filosóficos, es esencial insistir en este punto: las sensaciones de dolor son causadas por una serie de eventos que comienzan en terminales nerviosas libres y terminan en el tálamo y en otras regiones del cerebro. En lo que concierne a las sensaciones específicas, los eventos en el sistema nervioso central son, por cierto, suficientes para causar dolores, como sabemos por los dolores de miembros amputados [*phantom limb pains*][14] y por los dolores causados por porciones relevantes del cerebro artificialmente estimuladas. Y lo que es verdadero del dolor lo es también de los fenómenos mentales en general. Para decirlo con crudeza, y tomando al resto del sistema nervioso central como parte del cerebro a los fines de esta discusión: todo lo que importa en nuestra vida mental, todos nuestros pensamientos y sentimientos, están causados por procesos dentro del cerebro. En lo que concierne a la causación de los estados mentales, el paso crucial es el que sucede dentro de la cabeza y no el estímulo externo. Y el argumento en favor de esto es, simplemente, que si los eventos fuera del cerebro ocurrieran aunque sin causar nada en el cerebro, no habría eventos men-

14. Los dolores de miembros amputados son dolores que se sienten como proviniendo del ahora miembro inexistente.

tales, mientras que si ocurren eventos en el cerebro, los eventos mentales ocurrirían aun cuando no hubiera estímulo externo.

Creo que estos puntos son obvios, pero son inconsistentes con dos puntos de vista muy comunes acerca de la mente. Uno trata a la causación externa como el modo esencial de causación para los contenidos mentales. Pero en la explicación dada las cadenas causales externas sólo son importantes en la medida en que realmente impactan el sistema nervioso central. Otro punto de vista ampliamente sostenido es que no puede haber una relación causal entre los estados mentales y los cerebrales porque los estados mentales son estados cerebrales, y la forma de la relación de identidad en cuestión excluye la posibilidad de toda relación causal psicofísica. Por ejemplo, muchos filósofos materialistas solían afirmar que los dolores sólo *son* estimulaciones de las fibras C, mientras que según la explicación dada, las estimulaciones de las fibras C no son *idénticas* a los dolores, pero son *parte de las causas* de (ciertos tipos de) dolor.

Formulemos entonces la siguiente pregunta obvia: si los dolores y demás fenómenos mentales son causados por procesos neurofisiológicos, ¿qué son los fenómenos mismos? Bueno, en el caso de los dolores, son obviamente tipos de sensaciones no placenteras, pero esta respuesta nos deja insatisfechos porque no nos aclara, por así decir, cómo localizar los dolores y otros fenómenos mentales, en relación con el resto del mundo en que vivimos. ¿Cómo encajan los dolores en nuestra ontología general? Pienso, nuevamente, que la respuesta a esta pregunta es obvia, aunque tomará algún tiempo exponerla. A nuestra primera afirmación, esto es, que los dolores y demás fenómenos mentales son causados por procesos cerebrales, tenemos que agregar una segunda afirmación: *los dolores y demás fenómenos mentales son características del cerebro*.

Uno de los objetivos primordiales de esta discusión es mostrar cómo esas dos proposiciones pueden ser verdaderas *al mismo tiempo*. Ese par de tesis puede generar diferentes grados de perplejidad filosófica. En un cierto nivel podemos sentirnos perplejos respecto de cómo los fenómenos mentales y los físicos pueden estar en relación causal, cuando uno es una característica del otro. ¿No nos llevaría esto a la espantosa doctrina de la *causa sui*? Esto es, ¿no implicaría que la mente se causa a sí misma? En el fondo, gran parte de nuestra perplejidad proviene de un malentendido acerca de la naturaleza de la causación. Es tentador pensar que siempre que A causa a B tienen que haber dos eventos discretos, uno identificado como la causa y el otro identificado como el efecto; que toda la causación funciona sobre el modelo del relámpago que causa el

trueno. Si adoptamos este modelo basto de la causación estaremos tentados de pensar que las relaciones causales entre el cerebro y la mente nos fuerzan a aceptar algún tipo de dualismo; que eventos en un reino, el 'físico', causan eventos en otro reino, el 'mental'. Pero esto me parece un error. Y la manera de resolver el error es adoptar un concepto más sofisticado de causación. Para hacer esto, apartemos por un momento nuestra atención de las relaciones mente-cerebro con el fin de observar otro tipo de relaciones causales que se dan en la naturaleza.

Macro- y micro-propiedades

La distinción entre micro y macro-propiedades de los sistemas, es habitual en la física. Considérese, por ejemplo, el escritorio frente a mí, o el arroyo que fluye fuera de la ventana de mi oficina. Cada sistema está compuesto por micro-partículas y las micro-partículas tienen características en el nivel de las moléculas y de los átomos, así como en el de las partículas subatómicas. Pero cada sistema tiene también ciertas propiedades como la solidez en el caso de la mesa, o la fluidez en el caso del arroyo, que son macro-propiedades o propiedades de superficie [*surface properties*] de los sistemas físicos. Algunas macro-propiedades, pero no todas, pueden ser causalmente explicadas en base al comportamiento de los elementos en el micro-nivel. Por ejemplo, la solidez de la mesa frente a mí es explicada (causalmente) por la estructura reticular [*lattices*] de la que la mesa se compone. De manera similar, la fluidez del agua es explicada (causalmente) por el comportamiento de los movimientos de las moléculas de H_2O. Pero no todas las macro-propiedades tienen una explicación causal en términos de micro-comportamiento. Por ejemplo, la velocidad del arroyo no se explica en base al movimiento de las moléculas sino más bien por el ángulo de la pendiente, la atracción de la gravedad y la fricción provista por el lecho. Pero en el caso de las macro-características que son explicadas causalmente en base al comportamiento de los elementos en el micro-nivel, me parece que tenemos un modelo perfectamente normal para explicar las intrigantes relaciones entre la mente y el cerebro. En el caso de la solidez y la fluidez, no tenemos ninguna dificultad en decir que los fenómenos de superficie son *causados por* el comportamiento de elementos del micro-nivel y, al mismo tiempo, que los fenómenos de superficie *sólo son* rasgos (físicos) de los sistemas en cuestión. Mi modo preferido de enunciar este punto es decir que la característica de superficie F es *cau-*

sada por el comportamiento de los micro-elementos M, y que al mismo tiempo está *realizada en* [*realized in*] el sistema de los micro-elementos. Las relaciones entre F y M son causales pero al mismo tiempo F es, simplemente, una característica de nivel más elevado del sistema que consiste en los elementos M.

En contra de esto se podría decir que F sólo es idéntico a las características de M. Así, por ejemplo, podríamos definir la solidez como la estructura reticular del arreglo molecular [*molecular arrangement*]. Este punto me parece correcto, pero no considero que sea una objeción al análisis que propongo. Es típico del progreso de la ciencia que una expresión que es definida originariamente en términos de las características de superficie de un fenómeno, características accesibles a los sentidos, sea definida subsiguientemente en términos de la micro-estructura que las causa. Así, para tomar el ejemplo de la solidez, la mesa frente a mí es sólida en el sentido ordinario de que es rígida, que resiste cierta presión, que sostiene libros, que no es fácilmente penetrable por otros objetos tales como otras mesas, etcétera. Tal es la noción de sentido común de la solidez. Ahora bien, en vena científica uno puede definir la solidez como cualquier micro-estructura que cause esos toscos rasgos observables. Uno puede decir que la solidez es sólo la estructura reticular del sistema de moléculas y que la solidez así definida causa, por ejemplo, la resistencia al tacto y a la presión; o, uno puede decir que la solidez consiste de cosas tales como la rigidez y la resistencia al tacto y a la presión, y que está causada por el comportamiento de los elementos en el micro-nivel. Este paso de la causación a la identidad definicional es muy común en la historia de la ciencia. Considérese los siguientes pares: un relámpago es causado por una descarga eléctrica —el relámpago sólo es una descarga eléctrica; el color rojo es causado por emisiones de fotones con una longitud de onda de 600 nanómetros— rojo sólo es una emisión de fotones de 600 nanómetros; el calor es causado por movimientos de moléculas —el calor es sólo la energía cinética media de los movimientos de las moléculas.

Si aplicamos estas lecciones al estudio de la mente me parece que no hay dificultad en dar cuenta de las relaciones metafísicas de la mente con el cerebro en términos de una teoría causal del funcionamiento del cerebro como productor de estados mentales. Así como la fluidez del agua está causada por el comportamiento en el micro-nivel y es al mismo tiempo una característica realizada en el sistema de micro-elementos, exactamente en el mismo sentido de 'causado por' y 'realizado en', los fenómenos mentales son causados por los procesos que suceden

en el cerebro en el nivel neuronal o modular, pero están realizados en el mismo sistema que consiste en neuronas organizadas en módulos. Y así como necesitamos la distinción micro-macro para cualquier sistema físico, por las mismas razones necesitamos la distinción micro-macro para el cerebro. Aunque podemos decir de un sistema de partículas que es sólido o líquido, no podemos decir de una partícula dada que sea sólida o líquida. De la misma forma, hasta donde sabemos, aunque podemos decir de un cerebro particular que ese cerebro es conciente o que está experimentando sed o dolor, no podemos decir de una neurona en particular que sienta dolor o que experimente sed. Reiterando el punto una vez más: aunque hay misterios empíricos enormes acerca de cómo funciona en detalle el cerebro, no hay obstáculos lógicos, filosóficos o metafísicos en dar cuenta de la relación entre la mente y el cerebro en términos que nos son completamente familiares respecto del resto de la naturaleza. Nada es más común en la naturaleza que el que los rasgos de superficie de un fenómeno sean causados por una micro-estructura y realizados en ella; y ésas son, exactamente, las relaciones que son exhibidas por la relación de la mente con el cerebro. Las características intrínsecamente *mentales* del universo son las características *físicas* de alto nivel de los cerebros.

La posibilidad de los fenómenos mentales

Retornemos ahora a los cuatro problemas que parece enfrentar toda solución putativa del problema mente-cerebro.

¿Cómo es posible la conciencia?

La mejor forma para mostrar cómo algo es posible es mostrar cómo es en efecto, y ya hemos dado un esquema de cómo los dolores son causados efectivamente por los procesos neurofisiológicos que suceden en el tálamo y en el córtex sensorial. ¿Por qué es que mucha gente no se siente satisfecha con este tipo de respuesta? Creo que si se hace una analogía con un problema anterior en la historia de la ciencia podemos disipar esa sensación de perplejidad. Durante largo tiempo muchos biólogos y filósofos pensaron que era imposible, en principio, dar cuenta de los fenómenos de la vida apelando a un fundamento puramente biológico. Pensaron que además de los procesos biológicos era necesario otro elemento, tenía que postularse algún *élan vital* para dar vida a lo

que de otra manera era materia muerta e inerte. Es muy difícil darse cuenta hoy en día de lo intensa que fue la disputa entre el vitalismo y el mecanicismo, una generación atrás. Hoy esas cuestiones no se toman más en serio. ¿Por qué? ¿Es sólo porque hemos sintetizado la urea (el primer componente orgánico en ser sintetizado) y eso probó que los componentes orgánicos podían ser producidos artificialmente? Pienso que no. Pienso en cambio que es porque hemos llegado a visualizar el carácter biológico de los procesos que son típicos de los organismos vivientes. Una vez que comprendemos cómo las características que son típicas de los seres vivientes tienen una explicación biológica, deja de ser misterioso para nosotros que la materia inerte deba tener vida. Consideraciones exactamente análogas deberían aplicarse a nuestra discusión acerca de la conciencia. Que este trozo de materia inerte, esta sustancia gris y blanca con textura de harina de avena, sea conciente, no debería resultar más misterioso que lo que nos parece problemático que este trozo de materia, esta colección de ácidos nucleicos, proteínas y otras moléculas adosado a una estructura de calcio, esté viva. En síntesis, el modo de disipar el misterio es comprender los procesos. Todavía no entendemos los procesos completamente, pero entendemos el *carácter* de los procesos, entendemos que hay ciertos procesos electroquímicos que acaecen en las relaciones entre neuronas o entre módulos de neuronas y quizás otras características del cerebro, y que esos procesos son causalmente responsables de los fenómenos de conciencia.

¿Cómo pueden tener intencionalidad átomos en el vacío?

Como en el caso de nuestra primera pregunta, la mejor manera de mostrar cómo algo es posible es mostrar cómo es en efecto. Considérese la sed. Hasta donde sabemos, al menos ciertos tipos de sed son causadas en el hipotálamo por secuencias de disparos de neuronas. Estos disparos son a su vez causados por la acción de la hormona peptídica angiotensina II en el hipotálamo, y la angiotensina II, a su vez, es sintetizada por la renina, la cual es secretada por los riñones. La sed, al menos la de estos tipos, es causada por una serie de eventos en el sistema nervioso central, principalmente en el hipotálamo, y se realiza en el hipotálamo. Adviértase que la sed es un estado intencional. Tener sed es tener, entre otras cosas, deseo de beber. La sed tiene contenido proposicional, dirección de ajuste, condiciones de satisfacción y todos los otros rasgos que son comunes a los estados intencionales.

Como con los 'misterios' de la vida y de la conciencia, la manera de

aclarar el misterio de la intencionalidad es describir con todo el detalle que podamos cómo los fenómenos son causados por procesos biológicos, al mismo tiempo que se realizan en sistemas biológicos. Las experiencias visuales y auditivas, las sensaciones táctiles, de hambre, de sed, el deseo sexual y las experiencias olfatorias, son todas causadas por procesos cerebrales y se realizan en la estructura del cerebro, y todas son fenómenos intencionales. No estoy diciendo que debamos perder nuestra percepción de los misterios de la naturaleza; por el contrario, los ejemplos que cité son todos, en algún sentido, asombrosos. Pero quiero decir que ellos no son ni más ni menos misteriosos que otros asombrosos rasgos del mundo, como la existencia de la fuerza gravitacional, el proceso de fotosíntesis o el tamaño de la Vía Láctea.

Subjetividad

El enigma de la subjetividad puede enunciarse de manera bastante simple. Desde el siglo diecisiete nuestra concepción de la realidad ha involucrado la noción de objetividad total [*total objectivity*]. La realidad, según esa visión, es aquello que es accesible a todo observador competente. En algunas versiones, la realidad es lo que es medible objetivamente. Ahora bien, la pregunta es, ¿cómo acomodamos la subjetividad de los estados mentales en este cuadro?, ¿cómo encuadramos en una concepción objetiva del mundo real el hecho de que cada uno de nosotros tiene en realidad estados mentales subjetivos? La solución a este enigma también puede enunciarse simplemente. Es un error suponer que la definición de la realidad deba excluir la subjetividad. Si 'ciencia' es el nombre del conjunto de verdades objetivas y sistemáticas que podemos enunciar acerca del mundo, entonces la existencia de la subjetividad es un hecho científico tan objetivo como cualquier otro. Si una explicación científica del mundo intenta describir cómo son las cosas, entonces uno de los rasgos de la explicación será la subjetividad de los estados mentales, ya que es un hecho simple de la evolución biológica el que haya producido ciertos tipos de sistemas biológicos, a saber, el humano y ciertos cerebros de animales, que tienen rasgos subjetivos. Mi estado de conciencia actual es una característica de mi cerebro y en consecuencia me es accesible de una manera que no le es accesible a otro, y el estado de conciencia actual suyo es un rasgo de su cerebro que le es accesible a usted de una manera que no me es accesible a mí. Así, la existencia de la subjetividad es un hecho físico objetivo de la biología. Un error recurrente consiste en tratar de definir 'ciencia' en términos

de ciertas características de las teorías científicas existentes. Pero una vez que ese provincialismo es percibido como el prejuicio no científico que es, entonces cualquier dominio de hechos está sujeto a la investigación científica. Si, por ejemplo, Dios existiera, entonces ese hecho sería un hecho de la ciencia, como cualquier otro. No sé si Dios existe, pero no tengo duda de que los estados mentales subjetivos existen, porque yo tengo ahora uno y usted también. Si el hecho de la subjetividad va en contra de cierta definición de 'ciencia', entonces habría que abandonar la definición y no el hecho.

Causación intencional

Para nuestro propósito, el problema de la causación intencional es cómo dar cuenta de lo mental de modo de evitar el epifenomenalismo. ¿Cómo, por ejemplo, algo tan gaseoso y etéreo como el pensamiento podría dar origen a una acción? La respuesta es que los pensamientos no son gaseosos ni etéreos. Sus propiedades lógicas e intencionales están solidamente fundadas en sus propiedades causales en el cerebro. Los estados mentales pueden causar la conducta mediante el proceso causal ordinario, porque son estados físicos del cerebro. Ellos tienen un nivel elevado y un nivel bajo de descripción, y cada nivel es causalmente real.

Para ilustrar esas relaciones podemos usar nuevamente una analogía con la física. Considérese el martillar un clavo con un martillo. El martillo y el clavo tienen que tener un cierto grado de solidez. Los martillos hechos de algodón o de manteca serían muy poco útiles, y los martillos hechos de agua o de vapor no son martillos. La solidez es una propiedad causal real del martillo y no algo epifenoménico. Pero la solidez misma está causada por el comportamiento de las partículas en el micro-nivel y se realiza en el sistema de los micro-elementos. La existencia de dos niveles causales reales de descripción en el cerebro, un macro-nivel de procesos neurofisiológicos mentales y otro micro-nivel de procesos fisiológicos neuronales, es exactamente análogo a la existencia de los dos niveles causales reales en la descripción del martillo. La conciencia, por ejemplo, es una propiedad causal real del cerebro y no algo epifenoménico. Mi intento conciente de realizar una acción, tal como levantar mi brazo, causa el movimiento del brazo. En un nivel más elevado de descripción, mi intención de levantar el brazo tiene al movimiento de mi brazo como su condición de satisfacción y causa el movimiento del brazo. En un nivel más bajo de descripción, una serie de disparos de neuronas que se originan en el córtex causan la descarga del neurotrans-

misor acetilcolina en las placas neuromusculares ['*end plates*'] en donde los axones terminales de las neuronas motoras se conectan con las fibras musculares; esto a su vez causa una serie de cambios químicos que resultan en la contracción del músculo. En el caso de martillar un clavo también sucede que la misma secuencia de eventos tiene dos niveles de descripción que son causalmente reales, y el nivel elevado de rasgos causales es causado y realizado en la estructura de los elementos de nivel más bajo.

Categorías tradicionales

Hasta aquí me he resistido a usar el vocabulario tradicional de dualismo, monismo, fisicalismo, etcétera al intentar caracterizar la posición defendida en este capítulo. Sin embargo, puede ser útil ver cómo esos enfoques se relacionan con las categorías tradicionales. En una discusión de estos temas durante la conferencia sobre Filosofía de la Mente que tuvo lugar en la New York University, Hilary Putnam, de Harvard University, caracterizo el enfoque que presento aquí como (1) dualismo de propiedades, (2) emergentismo [*emergentism*], (3) superveniencia [*supervenience*]. Creo que si consideramos cada una de estas evaluaciones profundizaremos la comprensión de estas cuestiones.

Dualismo de propiedades

Si por 'dualismo de propiedades' se entiende, simplemente, el punto de vista de que el mundo contiene algunos rasgos físicos que son mentales —mi actual estado de conciencia, por ejemplo— y algunos rasgos físicos que son no-mentales —el peso de mi cerebro, por ejemplo—, entonces mi punto de vista puede describirse correctamente como un dualismo de propiedades. Creo sin embargo que hay algo profundamente engañoso en esta caracterización. 'Dualismo de propiedades' parece implicar que hay dos y sólo dos tipos de propiedades en el mundo, el físico y el mental, y ése no es de ninguna manera, el punto de vista que sostengo. Para mí, las propiedades mentales sólo son características físicas de alto nivel de ciertos sistemas físicos, en el mismo sentido en que la solidez o la fluidez son características físicas de alto nivel de ciertos sistemas físicos. Así, las propiedades mentales son propiedades físicas en el sentido en que la fluidez y la solidez son propiedades físicas. Me parece que este punto de vista es descripto correctamente no tanto

como dualismo de propiedades sino como polismo de propiedades [*property polyism*]. Esto es, que hay cantidades de tipos diferentes de propiedades de sistemas de alto nivel y que las propiedades mentales están entre ellas. Para decirlo de otro modo, de acuerdo con mi punto de vista las palabras 'mental' y 'físico' no son opuestas entre sí porque las propiedades mentales, interpretadas ingenuamente, sólo son una clase de propiedades físicas, y las propiedades físicas se oponen correctamente no a las propiedades mentales sino a rasgos tales como las propiedades lógicas y las propiedades éticas, por ejemplo.

Emergentismo

Respecto de la pregunta de si debemos o no concebir a las propiedades mentales como emergentes, valen consideraciones similares. Todo depende de lo que uno signifique por 'emergente'. Si vamos a pensar como emergente a una característica de alto nivel del sistema, como la solidez, la fluidez, etcétera, entonces en ese sentido creo que los estados de conciencia, la intencionalidad, la subjetividad, etcétera, son propiedades emergentes de ciertos sistemas biológicos. De hecho, si definimos a las propiedades emergentes de un sistema de elementos como las propiedades que pueden ser explicadas por el comportamiento de elementos individuales pero que no son propiedades de los elementos interpretados individualmente, entonces es una consecuencia trivial de mi punto de vista que las propiedades mentales son propiedades emergentes de los sistemas neurofisiológicos. Sin embargo, tradicionalmente, se ha considerado que emergentismo implica algo misterioso, que hay un proceso misterioso no físico que produce un tipo peculiar de propiedad. En pocas palabras, el emergentismo tiende a compartir los aspectos más misteriosos del dualismo, y en ese sentido niego que mi punto de vista pueda caracterizarse correctamente como emergentista. Si se considera que el emergentismo implica algo misterioso en la existencia de las propiedades emergentes, algo que yace más allá del alcance de las ciencias físicas o biológicas tal como son normalmente interpretadas, entonces me parece claro que las propiedades mentales no son emergentes en ese sentido.

Superveniencia

La doctrina de la superveniencia de lo mental en lo físico dice que no puede haber diferencias mentales sin las correspondientes diferencias

físicas: si en dos momentos diferentes un sistema está en dos estados mentales diferentes, entonces en esos dos momentos tiene que tener propiedades físicas diferentes. Este enfoque es una consecuencia de la tesis de que los fenómenos mentales son causados por el cerebro y realizados en él, porque si los efectos son diferentes, las causas tienen que ser diferentes. Me parece, por cierto, que es un mérito del enfoque que he adelantado aquí, que la superveniencia de lo mental sea simplemente un caso especial del principio general de la superveniencia de macro-propiedades en micro-propiedades. No hay nada especial o arbitrario o misterioso acerca de la superveniencia de lo mental en lo físico; es simplemente un caso más de la superveniencia de las propiedades físicas de alto nivel en las propiedades físicas de nivel más bajo. Si un recipiente con agua tiene hielo en cierto momento y líquido en otro momento, entonces tiene que haber una diferencia en el comportamiento de las micro-partículas que dé cuenta de la diferencia. De manera semejante, si yo quiero agua en un momento y luego no quiero agua, tiene que haber una diferencia en mi cerebro que dé cuenta de esta diferencia en mis estados mentales.

Consecuencias para la filosofía de la mente

Algunos conceptos mentales como, por ejemplo, *tener un dolor* o *creer* que tal y cual, denotan entidades que existen enteramente en la mente. Otras como *ver* o *conocer* también se refieren a fenómenos mentales pero requieren que se satisfagan condiciones adicionales para que el concepto sea aplicable. Así, por ejemplo, decir que X sabe que P implica algo más que X crea que P; implica, entre otras cosas, que P es verdadero, y que la verdad de P no puede ser, en general, algo que suceda únicamente en la mente de X. Decir que X ve P implica que X tiene una experiencia visual de cierto tipo, pero también implica que es el caso que P. Llamemos a los conceptos cuyas condiciones de verdad dependen sólo de lo que sucede en la mente 'conceptos mentales puros' ['*pure mental concepts*'], y llamemos a los conceptos mentales cuyas condiciones de verdad requieren fenómenos extra-mentales, 'conceptos mentales híbridos' ['*hybrid mental concepts*']. Ahora bien, dado que los conceptos mentales híbridos contienen por definición un componente mental, en la medida en que discutimos la naturaleza de la mente, podemos separar el componente mental y examinarlo separadamente. Para cada concepto mental híbrido hay un concepto mental puro que le

corresponde, que capta el componente mental puro del concepto híbrido. En lo que concierne a la mente, podemos confinar nuestra discusión a los conceptos mentales puros y a los estados mentales puros que son las denotaciones [*denotations*] de los conceptos mentales puros. Toda vez que un fenómeno mental está presente en la mente de un agente —por ejemplo, siente dolor, piensa en la filosofía o desea beber una cerveza fría— las condiciones causalmente suficientes para el fenómeno están enteramente en el cerebro. Y por cierto que la tesis de que los fenómenos mentales están causados por el cerebro y realizados en él tiene la consecuencia de que, para cualquier fenómeno mental, hay condiciones causalmente suficientes en el cerebro. Llamenos a este principio, *el principio de suficiencia neurofisiológica* [*the principle of neurophysiological sufficiency*]. Ahora bien, si este principio es verdadero, entonces muchas teorías corrientes en la filosofía de la mente se tornarán falsas, porque son inconsistentes con él. Por ejemplo, varios filósofos que siguen a Wittgenstein y a Heidegger han tratado de explicar la intencionalidad de los fenómenos mentales en términos de relaciones sociales. Pero, ¿cómo hemos de tomar esa explicación? Si la tomamos como afirmando que las relaciones sociales son necesarias para la vida mental o constitutivas de ella, entonces sabemos que tiene que ser falsa, porque las relaciones sociales son relevantes para la producción causal de la intencionalidad sólo si impactan en los cerebros de los agentes humanos; y los estados mentales efectivos, las creencias, los deseos, las esperanzas, los temores y el resto de ellos, tienen condiciones causalmente suficientes que son internas, enteramente, al sistema nervioso. Esto no implica negar que las relaciones sociales sean cruciales para la producción de muchas formas de intencionalidad, tal como, por ejemplo, el lenguaje. Los niños pueden aprender y usar un lenguaje sólo si están expuestos a otras personas que también usan lenguaje. Pero la tesis de que hay formas de intencionalidad que requieren una base social necesita ser reinterpretada de modo de que sea consistente con la afirmación de que la intencionalidad es un producto puramente interno de los procesos fisiológicos internos. Esos enfoques no son necesariamente inconsistentes; sólo pueden interpretarse como formas de describir aspectos diferentes del mismo fenómeno. El error consiste en suponer que las relaciones sociales puedan de una manera u otra reemplazar o substituir lo que sucede en el cerebro.

Una negación implícita más destacada del principio de suficiencia neurofisiológica proviene de la tradición construida en torno a la afirmación de Wittgenstein de que 'un proceso interno requiere un criterio

externo'.[15] Así, por ejemplo, Norman Malcolm ha tratado de dar un explicación no interna del soñar,[16] Elizabeth Anscombe ha tratado de explicar las intenciones en términos de conducta externa,[17] y Anthony Kenny ha tratado de explicar muchas emociones en términos de su escenario social [*social setting*] y de sus consecuencias conductuales.[18] Pero es difícil interpretar esos análisis de modo que sean consistentes con el principio de suficiencia neurofisiológica. Cualesquiera sean los demás rasgos que los sueños puedan poseer, son causados por procesos neurofisiológicos. Y lo mismo vale para las intenciones y las emociones, como el miedo y la angustia. Ahora bien, quizá podríamos interpretar los enfoques de Malcolm, de Anscombe y de Kenny como describiendo simplemente restricciones a la posesión de un *vocabulario* para discutir los fenómenos mentales. Y quizá podríamos interpretar la aseveración de Wittgenstein como la aseveración de que un *vocabulario* que corresponda a los procesos internos requiere criterios externos. Pero si tomamos esas aseveraciones como aseveraciones acerca de la *naturaleza* de los fenómenos mentales mismos —esto es, que uno no puede tener un sueño o una intención o estar enojado a menos que ciertas condiciones externas sean satisfechas, condiciones externas al cerebro—, entonces sabemos que esas tesis tienen que ser falsas debido al principio de suficiencia neurofisiológica. Lo que sucede en la cabeza tiene que ser causalmente suficiente para cualquier estado mental.

Y, por supuesto, la tradición wittgensteiniana es en sí misma parte de una tradición mayor que persigue un análisis conductista o cuasi-conductista de los conceptos mentales. Y una vez más, por el principio de suficiencia neurofisiológica sabemos que esos esfuerzos están condenados al fracaso. No podemos definir los fenómenos mentales en términos de sus manifestaciones conductuales, porque sabemos que siempre es posible experimentar los fenómenos, con independencia de tener alguna manifestación conductual.

15. Wittgenstein, Ludwig (1973): *Philosophical Investigations*, trad. G.E.M. Anscombe, Nueva York, Macmillan.

16. Malcolm, Norman (1959): *Dreaming*, Londres, Routledge & Kegan Paul.

17. Anscombe, G.E.M. (1963): *Intention*, Ithaca, Nueva York, Cornell Univ. Press.

18. Kenny, Anthony (1963): *Action, Emotion and Will*, Londres, Routledge & Kegan Paul.

Algunas conclusiones

Los objetivos más polémicos de este capítulo respecto de las mentes y los programas, pueden sintetizarse con rapidez. Para que la claridad sea total, enunciaré un conjunto de 'axiomas' y derivaré las conclusiones relevantes.

Axioma 1. Los cerebros causan a las mentes.

Ésta es, sencillamente, la enunciación cruda del hecho empírico de que procesos causales relevantes en el cerebro son suficientes para producir cualquier fenómeno mental. Es importante volver a enfatizar que en lo que hace a los fenómenos mentales puros, no hay ninguna conexión esencial entre los procesos causales internos que son suficientes para los fenómenos mentales y las relaciones causales de *input-output* del sistema total. En principio, podríamos vivir toda nuestra vida mental sin tener ninguno de los estímulos apropiados o ningún comportamiento externo normal.

Axioma 2. La sintaxis no es suficiente para la semántica.

Ésta es una verdad conceptual o lógica que articula la distinción entre el nivel de los símbolos formales y el nivel del significado.

Axioma 3. Las mentes tienen contenidos; específicamente, tienen contenidos intencionales o semánticos.

Axioma 4. Los programas son definidos formalmente o sintácticamente.

Ahora bien, de estos puntos obvios podemos derivar algunas conclusiones discutibles.

Conclusión 1. En sí misma, la instanciación de un programa nunca es suficiente para tener una mente (por los axiomas 2, 3 y 4).

Esta conclusión es suficiente por sí misma para refutar a la inteligencia artificial fuerte.

Conclusión 2. La manera en que el cerebro causa a las mentes no puede ser sólo por la instanciación de un programa (axioma 1 y conclusión 1).

Conclusión 3. Cualquier artefacto que tenga una mente tendría que tener poderes causales equivalentes (al menos) a los del cerebro (por axioma 1 trivialmente).

Conclusión 4. Para cualquier artefacto que tenga una mente, el programa por sí mismo no sería suficiente para proveerle tal mente. El artefacto tendría que tener poderes causales equivalentes al cerebro (por las conclusiones 1 y 3).

Quienquiera que desee cuestionar las tesis centrales nos debe una especificación precisa de los 'axiomas' y las derivaciones que cuestione.

TRADUCTORA: Florencia Luna.
REVISIÓN TÉCNICA: Eduardo Rabossi.